Компьютерная лингвистика и интеллектуальные технологии

по материалам ежегодной Международной конференции «Диалог» (2010)

Выпуск 9 (16)

Computational Linguistics and Intellectual Technologies

Papers from the Annual International Conference "Dialogue" (2010)

Issue 9 (16)

УДК 80/81; 004
ББК 81.1
К63

Программный комитет конференции выражает искреннюю благодарность
Российскому фонду фундаментальных исследований за финансовую поддержку,
грант № 10-01-06014-г

К63 **Компьютерная** лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 26–30 мая 2010 г.). Вып. 9 (16). — М.: Изд-во РГГУ, 2010.

ISBN XXX-X-XXXX-XXXX-X

Сборник включает 94 доклада международной конференции по компьютерной лингвистике и интеллектуальным технологиям «Диалог 2010», представляющих широкий спектр теоретических и прикладных исследований в области описания естественного языка, моделирования языковых процессов, создания практически применимых компьютерных лингвистических технологий.

Для специалистов в области теоретической и прикладной лингвистики и интеллектуальных технологий.

УДК 80/81; 004
ББК 81.1

ISBN XXX-X-XXXX-XXXX-X

Предисловие

Девятый выпуск ежегодника «Компьютерная лингвистика и интеллектуальные технологии» содержит материалы 16-й Международной конференции «Диалог»

В программу конференции было отобрано 94 доклада, охватывающие все традиционные направления конференции:

- Лингвистическая семантика и семантический анализ
- Формальные модели языка и их применение
- Теоретическая и компьютерная лексикография
- Создание и применение компьютерных лексических ресурсов
- Корпусная лингвистика. Создание, применение, оценка корпусов
- Интернет как лингвистический ресурс. Лингвистические технологии в Интернете
- Извлечение знаний из текстов
- Модели общения. Коммуникация, диалог и речевой акт
- Анализ и синтез речи
- Компьютерный анализ документов: реферирование, классификация, поиск
- Машинный перевод

«Диалог» не только является ведущей российской конференцией по **компьютерной лингвистике**, но и предлагает свое особое понимание содержания этого направления разработок и исследований, в соответствии с которым одинаково важны оба компонента: и инженерный, и лингвистический. «Диалог» традиционно ориентирован на идею, что успех в области автоматического анализа языка может быть достигнут лишь с использованием полноценных языковых моделей и качественных лингвистических ресурсов. Это определяет необычную для конференций по компьютерной лингвистике концентрацию теоретических докладов. Фактически в рамках «Диалога» проводятся — то разделяясь на секции, то объединяясь на пленарных заседаниях и круглых столах — сразу две конференции: теоретическая и прикладная. Следует подчеркнуть, что традиционный объект интереса на «Диалоге» — не только письменный текст, но и звучащая речь, коммуникативные стратегии, невербальные компоненты процесса общения.

Важным новым обстоятельством является то, что компьютерные технологии сегодня не только заимствуют у лингвистики ее наиболее эксплицитные результаты, но и оказывают серьезное влияние на сами лингвистические методы и постановку исследовательских задач. С каждым годом на «Диалоге» все интенсивнее обсуждается проблемы объективности языковых данных, на которых основываются лингвистические описания, и способы учета тех новых явлений, с которыми лингвист сталкивается в результате колоссального расширения доступного для изучения языкового пространства. Отсюда — то особое внимание, которое уделяется на конференции вопросам определения и фиксации языковой нормы.

По традиции Программный комитет предложил участникам тему — доминанту конференции. В этом году такой темой стало «Создание и применение компьютерных лексических ресурсов». Таким образом, в фокусе внимания «Диалога 2010» находятся проблемы создания и использования лексических баз данных разного типа: от компьютерных словарей и тезаурусов до лексико-семантических ресурсов типа WordNet, FrameNet и типологических баз данных.

Особенность этой темы, как и примыкающей к ней проблематики создания и использования корпусов, — универсальность: сегодня корпуса и лексические базы в Интернете — самые востребованные ресурсы как для лингвистов-исследователей, так и для прикладников.

Тематика «Диалога» очень широка, и этот сборник не может охватить всё: мы рекомендуем сайт конференции www.dialog-21.ru всем, кому интересны проблемы компьютерной обработки естественного языка в целом. На сайте можно ознакомиться и с условиями участия в конференции и публикации в этом ежегоднике. Там же представлены обширные электронные архивы «Диалога», включая тексты всех сборников прошлых лет.

Программный комитет «Диалога»
Редколлегия ежегодника «Компьютерная лингвистика и интеллектуальные технологии»

Организаторы

Ежегодная конференция Диалог проводится под патронажем Российского Фонда Фундаментальных Исследований при организационной поддержке компании ABBYY. Основными учредителями конференции являются:

- Институт лингвистики РГГУ
- Институт проблем информатики РАН
- Институт проблем передачи информации РАН
- Компания ABBYY
- Компания Яndex
- Рос НИИ искусственного интеллекта
- Филологический факультет МГУ

При поддержке Российской ассоциации искусственного интеллекта

Международный программный комитет

Нариньяни Александр Семёнович, *председатель*	РосНИИ искусственного интеллекта
Буате Кристиан	Гренобльский университет
Богуславский Игорь Михайлович	Институт проблем передачи информации
Гельбух Александр Феликсович	Национальный политехнический институт, Мехико
Зарецкая Елена Наумовна	Академия народного хозяйства при Правительстве РФ
Кибрик Александр Евгеньевич	Филологический факультет МГУ
Кобозева Ирина Михайловна	Филологический факультет МГУ
Козеренко Елена Борисовна	Институт проблем информатики РАН
Кронгауз Максим Анисимович	Институт лингвистики РГГУ
Лауфер Наталия Исаевна	ООО «проФан Продакшн»
Мельчук Игорь Александрович	Монреальский университет
Ниренбург Сергей	Университет Нью-Мексико
Осипов Геннадий Семёнович	Институт программных систем РАН
Попов Эдуард Викторович	РосНИИ информационной техники и систем автоматизации проектирования
Сегалович Илья Валентинович	Компания Яndex
Селегей Владимир Павлович	Компания ABBYY
Сулейманов Джавдет Шевкетович	Институт информатики КГУ
Флюр-Семёнова Вера	Компания SCIPER
Ыйм Халдур	Тартуский университет

Организационный комитет

Селегей Владимир Павлович, *председатель*	Компания ABBYY
Азарова Ирина Владимировна	Санкт-Петербургский государственный университет
Добров Борис Викторович	НИВЦ МГУ
Зацман Игорь Моисеевич	Институт проблем информатики РАН
Иомдин Леонид Лейбович	Институт проблем передачи информации РАН
Лауфер Наталия Исаевна	ООО «проФан Продакшн»
Лукашевич Наталья Валентиновна	НИВЦ МГУ
Перцов Николай Викторович	Институт русского языка им. В. В. Виноградова РАН
Соколова Елена Григорьевна	РосНИИ искусственного интеллекта
Толдова Светлана Юрьевна	Филологический факультет МГУ
Шаров Сергей Анатольевич	РосНИИ искусственного интеллекта

Секретариат

Левченкова Ирина Анатольевна, *секретарь оргкомитета, редактор сайта*	Компания ABBYY
Панфёрова Татьяна Витальевна, *координатор*	Компания ABBYY

Рецензенты

Азарова Ирина Владимировна
Апресян Валентина Юрьевна
Арефьев Андрей Александрович
Баранов Анатолий Николаевич
Беликов Владимир Иванович
Богуславский Игорь Михайлович
Браславский Павел Исаакович
Габрилович Евгений
Гайван Анна Александровна
Губин Максим
Добров Борис Викторович
Добровольский Дмитрий Олегович
Добрынин Владимир Юрьевич
Ермаков Александр Евгеньевич
Зарецкая Елена Наумовна
Захаров Леонид Михайлович
Иомдин Борис Леонидович
Иомдин Леонид Лейбович
Кибрик Андрей Александрович
Кобозева Ирина Михайловна
Козеренко Елена Борисовна
Крейдлин Григорий Ефимович
Кронгауз Максим Анисимович
Лахути Делир Гасемович
Левонтина Ирина Борисовна
Лобанов Борис Мефодьевич
Лукашевич Наталья Валентиновна
Ляшевская Ольга Николаевна
Масленников Мстислав
Некрестьянов Игорь Сергеевич
Ножов Игорь Михайлович
Осипов Геннадий Семёнович
Палажченко Павел Русланович
Панкратов Дмитрий Васильевич
Плунгян Владимир Александрович
Подлесская Вера Исааковна
Савельев Василий Евгеньевич
Сегалович Илья Валентинович
Селегей Владимир Павлович
Смирнов Иван Валентинович
Сокирко Алексей Викторович
Соколова Елена Григорьевна
Тестелец Яков Георгиевич
Тихомиров Илья Александрович
Толдова Светлана Юрьевна
Филатова Елена Александровна
Филиппова Екатерина
Циммерлинг Антон Владимирович
Янко Татьяна Евгеньевна

Содержание

Устное дистантное общение: о некоторых текстовых особенностях диалогов по мобильному телефону

Distant oral communication: some text characteristics of mobile phone dialogues

Алхимова И. С. (ioficerova@yandex.ru)

Институт русского языка РАН, Москва, Россия

Данная работа посвящена исследованию характерных особенностей диалогов по сотовому телефону. Анализ проводится на материале корпуса текстов, представляющих собой записи разговоров. Проанализированы некоторые характерные слова и реплики начала, середины, конца разговора.

Доступность сотовой связи и, как следствие, активное расширение круга пользователей и сферы применения этих телефонов привело к тому, что в нашем повседневном речевом общении появился новый вид коммуникации. Это обстоятельство породило и ряд собственно лингвистических проблем: в частности, формирование новых норм речевого поведения в ситуации общения по мобильному телефону; выработка новых речевых клише и стереотипов, специфичных для данной ситуации и не всегда совпадающих с ситуацией «обычного» телефонного общения, и т. д.

В данной работе будут проанализированы некоторые текстовые особенности разговоров по мобильному телефону.

Материалом для исследования послужил корпус из 70 аутентичных текстов, представлявших собой блокнотные и диктофонные записи разговоров по мобильному телефону. В основном, записи производились в тех ситуациях городского общения, которые позволяют непосредственно наблюдать и фиксировать подобные разговоры, — в транспорте, общественных учреждениях, на улице.

Следует учесть, что запись реальных диалогов имеет определенные трудности. Исследователь, являющийся внешним, сторонним наблюдателем речевого общения, может записать речь лишь одного из собеседников, реплик другого говорящего он не слышит и, соответственно, зафиксировать не может (за исключением тех случаев, когда в роли информанта выступал сам автор). Поэтому все записанные разговоры делились на две группы: исходящие (когда доступный для наблюдения и записи собеседник являлся инициатором разговора, или адресантом — 32 текста) и входящие (когда были зафиксированы ответные реплики того, кому позвонили, или адресата — 38 текстов). Кроме того, в процессе записи и при расшифровке небольшая часть информации может быть утеряна, но, как представляется, это не приводит к значительным изменениям результатов анализа.[1]

В качестве информантов выступали горожане, носители русской разговорной речи и городского просторечия. В большинстве случаев полную паспортизацию информанта получить нельзя, поэтому в качестве критерия отбора информантов использовалось их речевое поведение. По речевым особенностям (фонетическим, грамматическим и т. д.) можно определить некоторые социальные параметры говорящего, и речь людей с ярко выраженным иноязычным акцентом или диалектизмами, по возможности, не рассматривалась.

В целом разговор по сотовому телефону строится по модели «обычного» телефонного диалога, но имеет целый ряд текстовых особенностей, которые обусловлены спецификой самой ситуации «мобильного» общения: «разомкнутость» городского пространства, в котором находится абонент мобильной связи, и возможность перемещения в этом пространстве, присутствие при разговоре посторонних, становящихся невольными его слушателями, часто — незнание говорящим местонахождения собеседника (при общении по домашнему или служебному телефону такое исключено).

---

[1] Например, при блокнотной записи иногда трудно зафиксировать просодические особенности высказываний.

Композиционно любой диалог по мобильному телефону распадается на три части: начало разговора, основная часть и завершение разговора. Далее текстовые характеристики каждого из выделенных фрагментов будут рассмотрены отдельно. Нетрудно заметить, что в большинстве своем эти характеристики во многом те же, что и при разговоре по стационарному телефону. Однако исследования композиционно-текстового строения «обычных» телефонных разговоров сравнительно немногочисленны[2], поэтому мы вынуждены избрать менее экономное представление материала, т. е. давать описание всех встречающихся в данных текстах языковых явлений, обращая особое внимание на факты, характерные именно для данного типа коммуникации.

1. Начало разговора: приветствие

Прежде всего, следует отметить, каким образом собеседники вступают в разговор. Адресат чаще всего начинает диалог так же, как это происходит в разговоре по стационарному телефону — словом Алло (Алё), которое предваряет реплику-приветствие (если оно есть). Ср. (1) *-Алё// -… -Привет//*. Адресант чаще начинает разговор с приветствия (или сразу переходит к сути, опуская этикетные фразы); если же слово *Алло (Алё)* все же употребляется, то оно обычно не образует отдельной реплики. Ср.(2) *-… -Алё// Разбудила что ль? -…*;(3) *-…-Алё// Привет// -….*

Начало разговора и, в частности, приветствие в ситуации общения по мобильному телефону иногда отличается от приветствия в разговоре по стационарному телефону. Аппарат обычного телефона всегда находится в каком-либо помещении, и это обстоятельство часто маркируется в речи. Человек, который звонит по номеру стационарного телефона, в реплике приветствия упоминает место, в котором надеется найти адресата. Ср. типичное начало разговора по обычному телефону: *Здравствуйте, а Катя* ***дома****?* ; *Здравствуйте// Елена Александровна* ***в офисе/ на месте****?*. Очевидно, что для диалогов по мобильному телефону такие реплики не характерны в силу того, что владельцы сотовых мало ограничены территориально и в момент звонка могут находиться где угодно (за исключением случаев, подобных тем, когда офисный телефон — сотовый аппарат, постоянно находящийся в офисе и по функциям близкий к стационарному).

Приведенные примеры указывают еще на один внеязыковой фактор, который влияет на речевое поведение позвонившего: стационарный телефон — собственность всех членов семьи (или используется несколькими работниками офиса). Если абонент набирает номер стационарного телефона, то он, чаще всего, не знает наверняка, кто ответит на звонок[3]. Поэтому нередко встречаются высказывания типа Алё*/ здравствуйте//* ***А можно Катю?*** (Этот «зачин» представляет собой речевой шаблон, он настолько стереотипен, что позвонивший, скорее всего, начнет разговор именно так, если только он не знает наверняка, что на звонок ответила сама Катя). Если же человек звонит на чей-то мобильный телефон, то ожидает услышать голос его владельца. Говорящий приветствует адресата, иногда обращаясь к нему по имени. Стандартное приветствие в этом случае (4) — *Привет/* или *Привет/ Вань//*.

На начальном этапе развития сотовой связи почти ни у кого не было определителя номера, но в настоящий момент эта услуга доступна практически всем. Несмотря на то, что и стационарных телефонов, оснащенных определителем номера, становится все больше, в целом проблема идентификации собеседников более характерна для разговоров по стационарному телефону. Так, для разговоров по мобильному не столь характерны реплики, когда звонящий представляется, а адресат спрашивает *А кто это?* или *Маш/ это ты?*, как это иногда случается в разговоре по стационарному телефону. Взяв трубку и еще не услышав голоса собеседника, адресат знает, с кем будет говорить и сразу обращается к позвонившему по имени. Ср. первые реплики диалогов: (5); *-Алё Ир// Ой тьфу/ Жень// (смеется);*(6) — *Да па//*; (7)- *Привет бабушка//*; (8)-*Привет Сань//* и т. д.

В результате анализа собранного корпуса текстов было сделано следующее наблюдение. Приветствие в разговоре по мобильному телефону нередко является факультативным элементом и опускается (в частности, в корпусе реплики приветствия опущены в 38 текстах). Например, адресант может не поздороваться, если уже видел собеседника в этот день или разговаривал с ним. Ср. (9). *-…-Володь/ ну что? -… -Давай// (записывает телефон) 236-05-20// Так это же рабочий//*. В данном случае очевидно, что позвонивший уже разговаривал с адресатом: об этом свидетельствует вопрос *Ну что?*, который указывает здесь на то, что позвонивший ожидает выполнения просьбы, озвученной ранее. Характерны ситуации, когда позвонивший, поприветствовав собеседника, сразу же в своей первой реплике «переходит к делу»: таким образом, он в какой-то степени вынуждает адресата опустить приветствие и реагировать не на этикетную, а на содержательную часть высказывания. [4]

---

[2] См., например, работы [Schegloff: 1993], [Schegloff and Sacks: 1973].

[3] В данном случае мы не берем во внимание ситуацию, когда в доме или квартире живет только один человек, так как считаем их не столь характерными.

[4] Об иллокутивном вынуждении в структуре диалога см. [Баранов, Крейдлин: 1992, № 2].

Рассмотрим некоторые показательные реплики.

1. Со стороны позвонившего: (10) — *Привет/ ты уже дошел?*; (11)*Привет/ а ты где?*. Ответы типа *?Привет// Дошел//* или *?Привет// Я в метро (дома…)//* звучат менее естественно, а поэтому и менее частотны. Можно предположить, что ответные реплики на вопросы адресанта были примерно такими: (12) *Да/ (дошел)//* ; (13) *В метро// Уже еду//*
2. Со стороны отвечающего на звонок: (14)-*Алё// -… -Нормально/*; (15) — *Алё// -…- В электричке//*. Основываясь на этих репликах, можно реконструировать слова второго участника диалога: (16) *(Привет)/ как дела?*, (17) *(Привет)/ (а) ты где?*.

Очевидно, что примерные микродиалоги, которые можно построить из реконструированных и произнесенных реплик, будут схожи. Правильность восстановленных реплик подтверждается соответствующими произнесенными. Таким образом, формально реплика позвонившего состоит из приветствия и вопроса или только вопроса, а реплика адресата представляет собой ответ без приветствия.

Заметим, что «редукция» реплик приветствия обоими партнерами или приветствие в таком эллиптированном варианте, когда здоровается только один из собеседников, по-видимому, более характерна для целеориентированных, «деловых» диалогов. Ведь для фатического, нецелеориентированного общения соблюдение этикетных реплик более важно, так как последние выполняют первостепенную для такого общения функцию — установление контакта.

2. Основная часть разговора:

2.1. *«Ты где?»* — тема

Уже упоминалось, что говорящий по мобильному телефону “не привязан” к какому-то определенному месту общения. Само название *мобильный* определяется тем, что его владелец может взять телефон и отправиться куда угодно, его перемещения ничем не ограничены. Поэтому в начале разговора, сразу после приветствия или даже вместо него, часто задается вопрос *Ты где (сейчас)? (зафиксировано в 9 текстах исходящих звонков)*. Но в контексте информативного общения по мобильному телефону этот вопрос может иметь дополнительные элементы значения. Это связано с высокой конситуативной обусловленностью высказываний в разговорной речи[5].

Типовое высказывание «Ты где?» следует рассматривать как конситуативно обусловленное, определяемое ситуацией общения. В ряде контекстов употребление данного сочетания подразумевает заинтересованность говорящего в местонахождении другого участника диалога. Например, если собеседники договорились встретиться в каком-то месте, но по каким-либо причинам, не могут найти друг друга, выходом из такого положения становится звонок по мобильному и последующий вопрос *Ты где?*. Типичными в такой ситуации являются диалоги (18) *…- Ну вы где? — Мы возле дома гуляем// — Че-то я вас не вижу// — Ну… Ой//Всё// Посмотри на право// — А-а// Ну идите к подъезду//*.

Однако сочетание *ты где?* может содержать невербализованный смысл, подразумевающий не только заинтересованность говорящего в местонахождении собеседника, но и в том, чем он занят, чтобы помочь позвонившему скоординировать свои планы. Позвонивший, произнося фразу *Ты где?*, может интересоваться, например, тем, по каким причинам участника диалога в данный момент нет в условленном месте. Ср.: (19)(Девушка, в университете; спрашивает однокурсницу, приедет ли та на занятие, уже начавшееся) -… *-Ты где?* -… *-Ангиной?!* -… *-Понятно//* -… *-Давай// Пока//*.Ответ однокурсницы выглядел, вероятно, так *Дома// Я ангиной заболела//* Девушка поняла истинную цель позвонившей ей однокурсницы, которая на самом деле спрашивала: ‘Что случилось? Почему тебя нет на занятии?’ — и правильно отреагировала.

В обоих ситуациях шаблонный вопрос «Ты где?» отражает вполне конкретные и определенные интенции говорящего. Это вопрос, основная цель которого — спланировать или скоординировать действия собеседников.

Иногда вопрос *Ты где?* не является целеориентированным, а выступает в роли этикетной реплики, подобной высказываниям вида *Как дела?, Что делаешь?* или *У тебя все нормально?,* то есть берет на себя фатическую функцию.[6] Ср.:(20) -*Алё// -… -Ну что/ я поехала домой// -… — А мама там осталась// Я поехала за карточкой медицинского страхования// Потом ей отвезу// -…-Потом домой вернусь// А ты щас где?-…-Ага/ ну давай/ пока//*. В этом диалоге собеседница, как кажется, задает вопрос просто из любопытства, без какой-либо конкретной цели; ведь спрашивает она об этом в конце разговора, а не в начале, то есть информация о местонахождении собеседника для нее не первостепенна.[7]

---

[5] О конситуативных высказываниях, их типологии и особенностях текстового строения см. [РРР 1981; автор раздела — Е. Н. Ширяев], [Т. Г. Винокур: 2005].

[6] О фатической функции общения см. [Т.Г. Винокур: 1993], [Китайгородская М.В., Розанова Н.Н.: 1999].

[7] О значениях вопроса *Ты где?* на материале английского языка см. в статье Why people say where they are during mobile phone calls [Laurier: 2001].

2.2. Ситуация переспроса: функции *Алло*! и *Чего*?

В этом небольшом параграфе речь пойдет о речевом поведении говорящих в коммуникативной ситуации, когда они что-то не расслышали из-за помех на линии, шума и т. д. Данную ситуацию, конечно же, нельзя назвать специфичной для рассматриваемого вида общения: мы часто переспрашиваем собеседника и тогда, когда общаемся по стационарному телефону. Однако при речевом взаимодействии по мобильному телефону в условиях шумной городской среды, а также из-за несовершенства сотовой связи подобные «стрессовые» ситуации возникают весьма часто, поэтому о них их следует упомянуть. Типичной реакцией человека, не расслышавшего собеседника, является переспрос. Наиболее часто в таких случаях произносятся реплики типа *Что?*, *Чё?*, *Чево?, Что/Чё/Чево ты сказал?*. При переспросе из-за помех на линии возможен повтор реплик: (21)*-…-Алё/приве-ет//-…-****Чё?-…-Чё?-…-Чё****?-…-Соскучилась//-…-****Чё?-…-****Ага/ да/ уже// Я навязчива?*. Отметим, что «чё»-реплики весьма частотны особенно в молодежной разговорной речи, их функции в диалоге разнообразны. В зависимости от конситуации подобные реплики могут иметь разное значение, которое маркируется соответствующей интонацией. Например, выражение со словом «чё» может выражать удивление или непонимание:(22) *-Ну потому что у меня сёдня День рожде-ени-ия// (смеется)-…-Ты чё!? (смеется)-…-У меня сёдня//-…-Ты чё/Лёш?!* (интонация недоумения или удивления по поводу того факта, что собеседник не помнит дня рождения говорящего)

Междометие *Алло (Алё)*, как и слово *Чё?*, может употребляться в диалоге с различной интонацией и в нескольких значениях. Например, *Алло* в начале диалога обычно произносится с вопросительной интонацией и имеет контактоустанавливающую функцию. *Алло* в данном случае можно было бы заменить на *Я Вас/ тебя слушаю*. А в ситуации, когда возникают помехи или другие проблемы со связью, междометие *Алло* употребляется в функции переспроса, выражая другое значение: «я Вас/ тебя не слышу». При этом меняется и интонационное оформление: *алло* может произноситься с растяжкой гласного *Алё-о-о* или часто повторяться, если помехи не прекращаются. Ср.:(23)… -Отдыхаю// *-****Алё//****-…-А-а/ у меня тоже щас начались выходные//-…-Что значит какого числа? Нет// Выходные просто// В пятницу я не учусь/ суббота воскресенье//-…-Отдыхаю//-…-****Алё//-****…-А-а// Щас ты будешь провада… пропадать//-…-Чево ты говоришь?-…-Да эт я слышала уже//Я думала ты чё-то еще сказал/ другое//-…-А чё ты делаешь на выходных?-… -****Алё-о-у****// (Разговор прервался)*. Сигналом переспроса является и многократный повтор реплики: (24) — ***Алё-алё-алё****/ не слышу/ перезвони//*.

3. Окончание разговора: завершающие реплики

Типичными этикетными репликами прощания, как и в ситуации общения по стационарному телефону, являются *Пока//* или *(Ну) всё// Пока// (зафиксированы в 22 текстах)*. В официальных диалогах (то есть в диалогах между людьми, которых связывают официальные отношения; в разговорах малознакомых людей), в разговорах людей старшего возраста может употребляться более формальное *До свидания//*. Использование слова *всё* (чаще в неформальных диалогах) дополнительно указывает на завершение диалога: говорящий, тем самым, дает собеседнику понять, что он сообщил или получил всю нужную информацию.

Кроме реплик *Пока*, *Увидимся, До завтра*, традиционно выполнявших функцию завершения разговора и прощания, окончание диалогов часто маркирует слово частица *Давай* или сочетание *Давай/ пока// (Давай в данной функции зафиксировано в 23 текстах)*. Следует заметить, что слово *давай* может выступать в диалогах не только в функции репликипрощания. В РР оно нередко употребляется рядом с императивным глаголом в репликах-прескрипциях типа: *Ну давай/ иди//; Давай/ делай уроки//; Давай/ подавай на поскорее (на стол)//*. Дискурсивное слово *давай* может также замещать отсутствующий императивный глагол, «втягивая» в себя его конситуативное значение:(25) ***А.*** *Показать тебе письмо?* ***Б. Давай****//*; (Мать ребенку) *Ну давай (собирай портфель)/ в школу же опоздаешь//*. Ср. фрагмент телефонного разговора:… *-Володь/ ну что? -… -****Давай*** *(‘диктуй’)// (записывает телефон) 236-05-20// Так это де рабочий// ….* Слово *давай* может выступать в одном диалоге в обоих функциях — как императив-побуждение адресата и как завершающая реплика-прощание: (26) *(Девушка, в аудитории) -… -Алё дружок ты где? -… -Идешь от метро? Мы короче в триста шестнадцатой/ там где у нас Венг… [-… [-эта/ да// Кх-кх// Три вот//(смеется) -… -Что? -… -Да/ да/ конечно// -… -****Давай-давай*** (побуждение: ‘приходи скорее’)*// Мы ждем// -… -Давай* (побуждение)*// -… -Угу-Угу//* ***Давай****//* (прощание). Заметим, что сема побуждения к действию сохраняется в семантике слова *Давай* и тогда, когда оно употребляется в качестве завершающей реплики диалога.

В качестве варианта прощания можно привести следующие реплики: (27) –Ну ладно*// До связи// -… -Давай// До связи//*.Специфичное для общения по мобильному телефону прощание *до связи*, возможно, построено по аналогии с сочетанием *до встречи.*

Кроме того, в результате изучения диалогов было сделано еще одно наблюдение. Как и при общении по стационарному телефону, при «мобильном» общении собеседники иногда произносят завершающие диалог реплики, но потом продолжают

беседовать, а затем снова прощаются. Такую коммуникативную ситуацию можно было бы описать разговорным выражением «никак не могут распрощаться». Ср. весьма характерные для телефонных разговоров примеры многократного прощания: (28)…*-Харашо// **Ну давай**// -.. -Есть конечно/ да// -…-**Ну хорошо/ всё// Ну давай**// -… -Ага// **Ну пока**//*; (29)… *-**Ну всё**// -… -Да// Вы пото… Гене звоните// Потому что он сёдня на сутках// И завтра будет там с утра// Он вас встретит// -…-Ну вы доедете до «Локомотива»// Он выйдет и заберет вас//-…-**Ну всё**//-…-Ну и что/делать?-… -**Ну ладно**// Ну вы….-…-Ну да// Как-то так надо//-…-**Ну давайте**//-…-Вечером может я позвоню тогда вам еще//-…-Нет// Я вечером вам позвоню/ взяли вы билет не взяли// Едете/ не едете//-…-Ну попозже// может часов в десять/ в одиннадцать//-…-Всё?-…-**Ну всё тогда// Давайте**//* .

3.1. Завершение разговора без прощания. Функции слов *хорошо* и *ладно*

Разговоры, в которых нет реплик прощания (во всяком случае, со стороны того человека, чьи реплики зафиксированы), заканчиваются сразу после достижения участниками диалога их коммуникативных целей. Например: (30)(*мужчина в маршрутном такси*) –… — Да Ир// Алё// –… — (*молча отключает телефон)*; (31)(*женщина в маршрутном такси*) –… — Да// –… — Я уже еду к Первомайской//.

Функцию завершения разговора могут также выполнять слова *хорошо, ладно, всё.* Ср. пример: (32)*(Мужчина средних лет, едет в метро) -Да// -…- Хорошо/свяжусь//*. В данном контексте *хорошо* означает согласие с сообщенной адресантом информацией (в частности, с просьбой) и одновременно указывает на относительную смысловую завершенность предшествующей части диалога[8]; а здесь маркирует окончание всего диалога. Этикетные реплики отсутствуют, скорее всего, по нескольким причинам: 1)позвонивший торопится; 2) отвечающий на звонок находится в метро, где разговор может прерваться в любой момент. 3) адресант, возможно, еще встретится в этот день с позвонившим.

Сходная функция окончания разговора и у разговорной частицы *ладно* в следующем примере: (33) *(Девушка, в общежитии) -Алё// -… -Нормально// -… -Да// Мне надо книжку купить/ и вешалку/ а то уже скоро зима// -… -А ты на каком? -… -Ветку говори// Какая ветка? -… -Ну ладно// Я тогда подъеду//*. Интересен также следующий пример, где в первом и во втором случаях релятив *Хорошо//* выражает согласие со словами собеседника и обещание выполнить его просьбу. В конце разговора присутствуют обе рассматриваемые частицы, но к ним абонент прибавляет еще один сигнал, подчеркивающий завершение разговора (34.)– *Ну всё*: (*женщина с ребенком в маршрутном такси) –… — Алё// –… — Хорошо// –… Хорошо –…–Ладно/ хорошо// Ну всё//*. В этой ситуации содержание просьбы партнера коммуникации проясняется из последующего разговора женщины **Ж.** с сыном **А.**: ***А.*** *Это папа?* ***Ж.*** *Да//* ***А.*** *Чего он хотел?* ***Ж.*** *Мёду просил купить//*.

Итак, в статье были рассмотрены некоторые текстовые характеристики диалогов по сотовому телефону. В частности, анализ реплик приветствия выявил, что они иногда отличаются от соответствующих реплик при общении по стационарному телефону. Например, адресант, ожидая услышать голос владельца телефона, не будет говорить *Здравствуйте/ А можно Х?,* он сразу приветствует адресата, обращаясь к нему по имени.

Особенностью основной части диалога по мобильному телефону является нередкое употребление вопроса *Ты где?*, под которым может подразумеваться заинтересованность говорящего не только в местонахождении другого участника разговора, но и, к примеру, в том, что собеседник делает в данный момент. В качестве показателей окончания разговора, помимо традиционных *пока* и *до свидания*, характерны слова *давай, хорошо и ладно*.

В заключение следует добавить, что данная статья представляет собой лишь начальный этап исследования диалогического общения по мобильному телефону и, безусловно, содержит скорее наблюдения, чем лингвистически выверенные выводы. В дальнейшем, работа будет направлена на обнаружение других особенностей данного вида общения, более детальные описание и исследование ряда характеристик диалогов (в частности, распределения сегментных и просодических единиц), а также сравнительный анализ коммуникации по городскому и мобильному телефонам.

---

[8] См. [Баранов, Крейдлин: 1992, № 3: 87].

Литература

1. *Баранов А. Н., Крейдлин Г. Е.* Иллокутивное вынуждение в структуре диалога // Вопросы языкознания, 1992. №2. С. 84–99.
2. *Баранов А. Н., Крейдлин Г. Е.* Структура диалогического текста: лексические показатели минимальных диалогов // Вопросы языкознания, 1992, №3. С. 84–93
3. *Винокур Т. Г.* Говорящий и слушающий: Варианты речевого поведения. М.: КомКнига, 2005.
4. *Винокур Т. Г.* Информативная и фатическая речь как обнаружение разных коммуникативных намерений говорящего и слушающего // Русский язык в его функционировании: Коммуникативно-прагматический аспект. М.: 1993.
5. *Земская Е. А., Китайгородская М. В., Ширяев Е. Н.* Русская разговорная речь: Общие вопросы. Словообразование. Синтаксис. М.: 1981.
6. *Китайгородская М. В., Розанова Н. Н.* Речь москвичей: Коммуникативно-культурологический аспект. М.: Русские словари, 1999.
7. *Laurier E.* Why people say where they are during mobile phone calls (abstract) // Society and Space, 2001. [Электрон.ресурс] — Электрон. дан. — Режим доступа: http://www.receiver.vodafone.com/07/articles/07_page01.html
8. *Schegloff E. A., Sacks H.* Opening up Closings. // Semiotica 8(4), 1973. P. 289–327 [Электрон.ресурс] — Электрон. дан. — Режим доступа: http://www.sscnet.ucla.edu/soc/faculty/schegloff/
9. *Schegloff E. A.* Telephone Conversation // Encyclopedia of Language and Linguistics, vol. 9, 1993. P. 4547–4549 [Электрон.ресурс] — Электрон. дан. — Режим доступа: http://www.sscnet.ucla.edu/soc/faculty/schegloff/

Именные модели управления с точки зрения Грамматики Конструкций*

Nominal case patterns from the viewpoint of Construction Grammar

Антошина С. А.
Мурманский государственный педагогический университет

Ляшевская О. Н. (olesar@mail.ru)
University of Tromsø (Тромсе, Норвегия)

Современные описания именного управления так или иначе отталкиваются от гипотезы наследования моделей управления в рамках словообразовательного гнезда (ср. *ударить молотком — удар молотком*, *ударить по столу — удар по столу*). В статье исследуются случаи появления новых форм управления у имен существительных и прилагательных по сравнению с их мотивирующими коррелятами: ср. *заслон порнографии; комитет, аналогичный нобелевскому* и др., а также проявление тех же свойств в конструкциях с support verbs, ср. *поставить заслон*. Рассматриваются две гипотезы: о наличии у имени собственной модели (Grimshaw 1990) и о том, что модель может быть присуща всей конструкции целиком.

1. Введение

Есть ли у существительного *заслон* собственная модель управления? Согласно традиционной точке зрения, чтобы ответить на этот вопрос, мы должны сравнить модели управления имени и соответствующего глагола. В общем случае, существует почти полный изоморфизм между глаголом и именем в том, что касается набора семантических ролей и способа их поверхностно-синтаксического оформления. Упрощенная картина показывает нам, что именительному и винительному падежам, оформляющим субъект и объект при глаголе, будет соответствовать родительный падеж у имен, косвенные падежи и предложно-падежные группы в обоих случаях остаются те же самые, ср. *разгружать вагоны — разгрузка вагонов, ухаживать за больными — уход за больными*. Имя, как и глагол, может привлекать творительный падеж для оформления субъекта (*взятие русской армией Азова*), но также может использовать чисто именные способы, согласование с притяжательным местоимением или прилагательным (*твое/рихтеровское исполнение*, примеры из Падучева 2009)[1]. РГ 1980 приводит случаи, когда у имен редуцируется вариативность оформления, присущая глаголам, ср. *поступить работать* и *на работу — поступление на работу*; и лишь в отдельных группах имен наблюдается изменение в способе оформления участников: *любить кого-н. — любовь к кому-н., приветствовать кого-н. — приветствие кому-н., сочувствовать кому-н. — сочувствие к кому-н., радоваться чему-н. — радость по поводу чего-н., восторгаться кем-чем-н. — восторг перед кем-чем-н., начать работать — начало работы*.

Валентности прилагательных также «естественно сравнивать с валентностями глагола» (Ку-

* Исследование осуществлено в рамках проекта «Процессы словообразования в текстовой динамике» Программы ОИФН РАН «Текст во взаимодействии с социокультурной средой: уровни историко-литературной и лингвистической интерпретации».

[1] Ср. здесь также переоформление зависимости между предикатом и актантом в конструкции «имя + причастие»: ср. *присылка оттиска — благодарю за присланный оттиск* (примеры из Падучева 1980).

стова 2006: 323). В целом, за исключением одного участника, выражаемого существительным, от которого прилагательное синтаксически зависит (*похожий голос*), при прилагательных участники выражаются значительно реже, чем при глаголе. Однако в случае, когда они все-таки находят поверхностное выражение, способ их оформления как правило совпадает с глагольным, ср. *походить на кого-н. в чем-н. — похожий на кого-н. в чем-н.* Тут также наблюдаются отдельные зоны расхождений, например, *благодарить кого-н. — благодарный кому-н., доверять кому-н. — доверчивый к кому-н., устоять против чего-л. — устойчивый к чему-н.*, но, в отличие от существительных, вариативность поверхностного оформления участников при прилагательных может быть выше, чем при глаголах, ср. *враждовать с кем-н. — враждебный кому-н.* и *по отношению к кому-н.* (примеры из РГ 1980).

Таким образом, определенный изоморфизм между оформлением участников при глаголах и именах заставляет исследователей видеть в модели управления глагола точку отсчета для описания (см. обсуждение этой проблемы в Сердовольсккая 2004, Пазельская 2009аб). С этих позиций, собственной моделью управления обладают имена, у которых модель отличается от глагольной, и имена, не производные от глагола (ср. *мир с Турцией*). Оппонентами «теории наследования» выступают Дж. Гримшо (Grimshaw 1990) и ее сторонники, которые утверждают, что любое имя события (не результата) обладает собственной аргументной структурой. Еще дальше идут лексикографы, беспристрастно фиксируя особенности управления у любых имен, в том числе неситуативных, и даже у отдельных значений имен (ср. МАС, ТКС, НОСС, FrameNet и др.).

В данной статье мы рассмотрим случаи появления новых форм управления у имен существительных и прилагательных на примере имен *заслон* и *аналогичный*. Нас будет интересовать вопрос, насколько их модели управления, отличающиеся от моделей управления мотивирующих слов, можно считать «собственными». А именно, на материале НКРЯ (Основной корпус) мы хотим проанализировать контекстное окружение имен и проследить, насколько оно мотивирует инновационную модель управления.

2. Имя *заслон*: значения и конструкции

Имя *заслон* является дериватом от глагола *заслонить/заслонять*(*ся*). Модель управления этого глагола (*кто* заслоняет *кого/что от чего* и *чем*) предполагает выражение следующих участников: Агенса (Snom), Экспериента/Пациенса (Sacc), Воздействия (от + Sgen) и Средства (Sins). Примеры, приводимые в толковых словарях для существительного *заслон*, свидетельствуют о том, что предложное управление сохраняется только в первом значении — физической преграды (ср. *заслон от ветра, от дождя, от движущихся песков*). Роль Средства уже не может быть выражена творительным падежом, однако возможна конструкция *заслон из* + Sgen (в случае если Средство — вещественный или множественный участник). Для выражения Агенса и Экспериента/Пациенса, согласно стандартному правилу, остается конструкция с родительным падежом.

Однако, если мы посмотрим на другое, метафорическое употребление существительного заслон:

> перен., кому (чему). Противодействие, препятствие. З. браконьерам. З. расточительству. [Словарь Ожегова-Шведовой][2]

то мы увидим, что меняется способ выражения участника Воздействие — на дательный падеж. Помимо этого, участник может выражаться также некоторыми предложно-падежными группами (*против чего-л., на пути чего-л., для чего-л., перед чем-л.* и *между чем.- л.* и др.). Как и почему это происходит? Попробуем ответить на этот вопрос с помощью корпусного исследования.

2.1. Структура значений

Всего в корпусе обнаружено 408 употребления имени, из них 9 % приходится на первое значение (артефакт, который создают или используют для защиты от чего-л. неприятного — ветра, волн, вражеских стрел и т. д., или физический объект, который заслоняет, не дает увидеть некоторый другой объект); 54 % — на второе значение (группа военных, прикрывающая отход войск; в последнее время *заслон* употребляется также для обозначения группы милиционеров или других лиц, препятствующих проходу людей во время демонстрации); 36 % — на рассматриваемое третье, абстрактное значение, являющееся метафорическим переносом от первого (примерами здесь могут служить закон как *заслон* наркотрафику, психологические силы внутри человека как *заслон* преступным намерениям, бюрократические *заслоны* на пути приватизации и т. д.). Минимальное число контекстов представляет имя *заслон* в значении ‘печная заслонка’ (13 употреблений, 2 %) и в сленговом выражении *делать заслон* (действие, служащее для маскировки воровства или другого мошенничества) (2 употребления).

---

[2] Помимо двух указанных значений, у имени *заслон* есть еще несколько: отряд военных или военной техники, предотвращающий проход врага, печная заслонка, прием в баскетболе, которые в данной статье остаются вне зоны нашего внимания.

2.2. Фрейм

Круг типичных ситуаций, ассоциируемых с именем *заслон*, (фрейм) включает следующих участников:

Заслон (физический или абстрактный) защищает Экспериента/Пациенса от (Негативного) Воздействия*

- Они [леса] являются заслоном *от юго-восточных сухих ветров*{Воздействие}.
- Корпус танка иногда может быть использован в качестве заслона *амбразуры ДЗОТ*{Пациенс}.

*В перцептивных ситуациях *Заслон* не дает Эксперианту увидеть Перцепт (аналог Воздействия)

- Прямо под окном, в сотне шагов, за бетонным забором, за густым заслоном деревьев виднелась *голубизна воды*{Перцепт}.

Агенс создает *Заслон*, из Средства — или Агенс может использовать Средство в качестве *Заслона*

- *Она [церковь]* {Агенс} не сумела поставить заслон разрушительным идейным тенденциям и нравственному беспределу.
- По границам участков создаются противопожарные заслоны *из лиственных пород*{Средство}.
- *Люди Синебродова*{Агенс} расположились так предусмотрительно, использовав *скамейку*{Средство} в качестве естественного заслона, что уйти, не заставив их убраться с дороги, было невозможно.

Контрагент преодолевает *Заслон*

- Итак, на пути к матчу на первенство мира *Фишеру*{Контрагент} осталось преодолеть последний заслон.

В ряде случаев важно Место, в котором находится *Заслон*. Местонахождение Экспериента/Пациенса и направление Воздействия предполагается заранее известным.

- Благодаря их мужеству, отваге и чувству долга серьезный заслон наркоугрозе поставлен *на таджикско-афганской границе*{Место}.

2.3. Управление

Если обратиться к метафорическим употреблениям имени *заслон*, которые, как мы видели, являются для него наиболее распространенными, то можно увидеть, что более чем в половине случаев в контексте существительного присутствует группа, выражающая участника Негативное Воздействие. Это может быть группа в дательном падеже (*заслон пошлости*, 45[3]), предложные группы *на пути + Sgen* (*заслон на пути мигрантов*, 15), *от + Sgen* (*заслон от неловких соболезнваний*, 9), *против + Sgen* (*заслон против вольнодумства*, 5), *для + Sgen* (*заслон для правительственной политики контрреформ*, 5), *между + Sins и Sins* (*заслон между я и миром*, 2), *перед + Sins* (*заслон перед мигрантами*, 2) и, наконец, адъективная группа (*миграционный заслон,* 1).

Итак, наблюдается значительное разнообразие способов выражения участника, причем большинство из них являются инновациями по сравнению с моделью управления глагола *заслонять(ся)*.

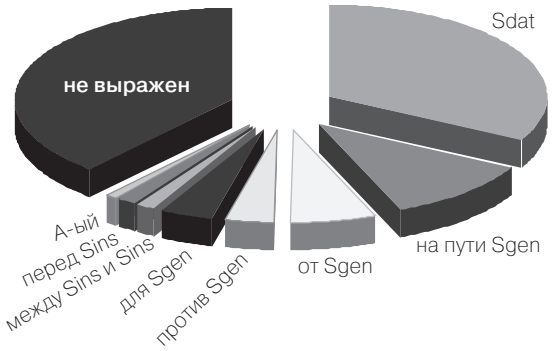


Диаграмма 1. Выражение участника Воздействие у имени ***заслон*** в метафорическом значении.

Для имени *заслон* наиболее характерны глаголы со значением создания (*соорудить, строить, городить* и др.), тождества/функции (*служить, выполнять роль, представлять собой*) и преодоления (*преодолеть, пробиться, продраться, просочиться, протолкнуться* и др.), но абсолютным чемпионом является глагол *ставить* и его производные (*поставить, выставить*; эти глаголы широко используются также при имени *заслон* в других значениях). В сочетании *ставить заслон* глагол также принадлежит к классу создания, однако его прототипическое значение — помещение объекта в пространстве. С этим фактом хорошо согласуется выбор четырех падежных групп со значением места (*на пути + Sgen, против + Sgen, перед + Sgen, между + Sins и Sins*):

1. Офицеры аппарата главы военного ведомства начали *ставить* изощренные бюрократические *заслоны на пути* тех, кто пытался пробиться на аудиенцию к маршалу.
2. Поездки, наука, лихорадочные метания из одной экзотической области знаний в другую — все было из-за этого, только книги не могли *выставить против* него роковой *заслон* и нехотя, с презрением открывались.
3. Юра думал, что этого не будет больше никогда, что, уехав на Сахалин, оборвав все нити, связывающие его с прежней жизнью, с Москвой, он *поставил* надежный *заслон перед* этим котлом, в котором как варево кипят отношения, для него невозможные.

[3] скобках приводится контекст из корпуса и количество найденных примеров.

4. Она [Гурченко] много натерпелась в жизни и частенько *ставит* эдакий *заслон между* собой и человеком, которого не хочет пустить в свой внутренний мир.[4]

Однако, чем же мотивирован дательный падеж? Этот способ, если и присутствует в моделях управления *ставить*, то для обозначения бенефицианта действия (ср. *ставить банки больному, поставить отцу телефон, нам поставили нового директора*) — то есть оформляет одушевленного участника.

С другой стороны, дативный участник характерен для таких глаголов как *препятствовать* (*нововведениям*, наряду с *в* + Sloc — *в нововведениях*), *преграждать* (*путь кому-л.*), *противодействовать, мешать, вредить* (с общим значением помехи), а также для их антонимов *способствовать, помогать, содействовать* (с пресуппозицией помехи). Ср. здесь же пару *запретить — разрешить* (*гражданам* выезд за границу).

Заметим также, что в конструкциях с дательным падежом можно наблюдать существительные *преграда, препятствие, помеха, барьер, заграждение* — они обозначают то, что может преграждать путь, и являются аналогами имени *заслон* как по значению, так и по сочетаемости; и существительные *доступ, путь* (ср. *открыть/закрыть доступ студентам/всему новому, преградить путь кому-чему-л.*). Наконец, имя традиционного бытового ограничителя пространства — *дверь* — также употребляются с дативом, но только в ситуации устранения помехи (ср. *открыть дверь врачу/вечности*), ср. Кузнецова, Ляшевская 2010.

Метафорически обобщенный фрейм обозначаемых ситуаций можно представить следующим образом: в ситуации присутствует (дативный) участник, который находится на пути к своей цели; агенс создает или устраняет препятствие на его пути.

Таким образом, можно утверждать, что дательный падеж в конструкции ставить заслон чему-л. не мотивирован ни управлением имени, ни управлением глагола, но присущ всей конструкции целиком.

3. Имя *аналогичный* и его подобия

Имя *аналогичный* не соотносимо с глаголом, однако оно имеет два родственных имени существительных, *аналог* и *аналогия*. Сравнение управления трех имен дает нам следующую картину: прилагательное *аналогичный* и имя *аналог* предпочитают для оформления контрагента (т. е. того, с кем сравнивают), дативную конструкцию, в то время как имя *аналогия* выступает вместе с союзом *с* и творительным падежом:

аналогичный — чему-л. *аналог — чему-л.*
аналогия — с чем-л.

В корпусе можно найти примеры, в которых *аналогичный* и *аналогия* идут друг другу навстречу и *аналогичный употребляется* в предложной конструкции, а аналогия — в дативной, но такие примеры очень редки:

1. Отношение автора к различным явлениям литературы и культуры носит диалогический характер, *аналогичный со взаимоотношениями* между хронотопами внутри произведения (о которых мы сказали выше). [М. Бахтин].
2. Условия службы персонала Организации, насколько это возможно, должны быть *аналогичны с условиями* в других Организациях Объединенных Наций.
3. В характеристиках этой таинственной сущности в этой статье есть удивительная *аналогия проблематике* понимания Пушкина, как она раскрывалась в русской мысли.
4. Сидящий на пятках прямыми руками давит на голову стоящему на четвереньках партнеру, *по аналогии схеме*, описанной выше.

В остальной же массе употреблений способ оформления прилагательного и существительного аналогия различаются. Вообще, мы можем сформулировать следующее правило: если некоторое явление характерно для конструкции А, то в языке должны найтись сходные конструкции, в которых наблюдается то же явление.

аналогичный чему	аналогия с чем аналог чему	*
подобный чему	подобие чему	подобать чему
тождественный чему	тождество с чем	тождествовать*
равный чему	равенство с чем	равняться чему
равный в чем	равенство в чем	равняться в чем
сходный с чем.	сходство с чем	сходиться с чем
сходный в чем	сходство в чем	сходиться в чем
схожий с чем		
схожий в чем		
похожий на что		походить на что
похожий в чем		походить в чем
похожий чем		походить чем
одинаковый с чем	*	
одинаковый в чем		
вылитый *	*	
конгениальный кому	конгениальность в чем	
разный в чем	разница в чем	разниться в чем

Следуя этому правилу, можно отметить параллелизм моделей у прилагательных *аналогичный, подобный, тождественный, равный* (чему-л.) и параллелизм предложных моделей у существительных *аналогия, тождество, равенство, сходство* (с чем-л.). Однако при переходе от одной части речи к другой способ оформления сохраняется только в цепочке *подобный — подобие — подобать* (кому-чему-л.) и *сходный/схожий — сходство — сходиться* (с кем-чем-л.), и именно эти цепочки (с хорошей внутренней формой, объясняющей выбор падежа и предлога) и можно считать мотивирующими для дативного падежа при *аналогичный* и предлога *с* при имени *аналогия*.

[4] Примеры из НКРЯ, Основной корпус.

4. Обсуждение

Итак, мы рассмотрели несколько случаев, когда модель управления слова не может считаться его собственной: если слово является производным и полностью наследует управляющие свойства исходного слова, если слово заимствует модель управления у слов, схожих синтаксически и семантически, и если оно приобретает те или иные управляющие свойства только в пределах особой лексико-грамматической конструкции. Как можно интерпретировать мотивированность поверхностно-синтаксического оформления участников. с точки зрения Грамматики Конструкций (ГК)? В этой теории конструкция постулируется как новая в том случае, если свойства целого не вытекают из свойств составлящих, ср. следующее рассуждение относительно статуса конструкции «только и делает/знает/умеет, что V»:

«Употребление глаголов *делать, знать, уметь* в конструкции также отличается от их независимого употребления. Для всех глаголов в конструкции характерна утрата различных ограничений на зависимый инфинитив, что является индикатором сдвига в значении глаголов. Особенно заметны отличия глагола *знать* в конструкции: полностью отсутствует значение, характерное для независимого употребления глагола; глагол, следующий после элемента что, не является зависимым, наследованным от исходной модели управления глагола, и обусловлен, по-видимому, влиянием двух других глагольных конструкций» (Овсянникова 2009).

В подходе, предлагаемом Grimshaw, Mester 1988, доказывалось, что свойства конструкции с light verbs получаются по определенным правилам из свойств конструкции вспомогательного глагола и семантически наполненного существительного[5]. В свете ГК такая полная детерминированность оформлении я ролей означала бы, что сочетания с light verbs новой конструкции не образуют.

В рассмотренных нами выше примерах понятие мотивированности поверхностного синтаксиса конструкции (Ляшевская, Рахилина 2008) не позволяет свести свойства конструкции к свойствам составляющих. В *поставить заслон пьянству* датив не свойственен ни модели управления имени *заслон*, ни модели вспомогательного глагола. В случае с прилагательным *аналогичный* дативная модель является общей для целого класса имен, то есть опять же не является собственной моделью прилагательного *аналогичный* — но уже в другом смысле. Скорее, наоборот, видя распространение модели управления у близких по смыслу слов одного частеречного класса, удивительно требовать одинаковой модели управлений у прилагательного, существительного и глагола.

---

[5] Впрочем, в той же статье Гримшо замечает: Grimshaw: конструкции с light verbs могут отличаться от конструкций с глаголами лексическим заполнением актанта: “although a spider can *walk*, a spider does not normally *take a walk*” (Grimshaw, Mester 1988: 229)

Литература

1. *Кузнецова Ю. Л., Ляшевская О. Н.* (2010). Конструкции и трансформации // Международная конференция «Слово и язык». Москва, 2–4 февраля 2010 г.
2. *Кустова Г. И.* (2006). Валентности и конструкции прилагательных // Лауфер Н. И., Нариньяни А. С., Селегей В. П. (ред.), Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2006». М.: РГГУ. С. 323–238.
3. *Ляшевская О. Н., Рахилина Е. В.* (2008). Многозначность сквозь призму дискурса // А. Мустайоки, М. В. Копотев, Л. А. Бирюлин, Е. Ю. Протасова (ред.), Инструментарий русистики: корпусные подходы. Slavica Helsingiensia Series. Вып. 34. Helsinki. С. 217–232.
4. *Овсянникова М.* (2009). Наследование и изменение свойств на различных уровнях конструкции: конструкции только и делает, знает, умеет, что V // Круглый стол «Русский язык: конструкционные и лексико-семантические подходы» (в рамках XXXVIII Международной филологической конференции). http://iling.spb.ru/nord/materia/rusconstr2009/ovsjannikova.pdf
5. *Падучева Е. В.* (1980). Об атрибутивном стяжении подчиненной предикации в русском языке // Машинный перевод и прикладная лингвистика. Вып. 20. С. 3–44.
6. *Падучева Е. В.* (2009). Посессивы и имена способа действия // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2009» (Бекасово, 27–31 мая 2009 г.). Вып. 8 (15). М.: РГГУ. С. 365–372.
7. *Пазельская А. Г.* (2009а). Модели деривации и синтаксическая позиция отглагольных существительных по корпусным данным // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2009» (Бекасово, 27–31 мая 2009 г.). Вып. 8 (15). М.: РГГУ.
8. *Пазельская А. Г.* (2009б). Модели деривации отглагольных существительных: взгляд из корпуса. // Киселёва К. Л., Плунгян В. А., Рахилина Е. В., Татевосов С. Г. (ред.) Корпусные исследования по русской грамматике. Москва.
9. *РГ 1980* — Русская грамматика. (1980). М.: Наука.
10. *Сердобольская Н. В.* (2004). Оформление актантов в номинализации и семантические классы глаголов. // LENCA-2. Казань, 11–14 мая 2004. с. 206.
11. *Grimshaw J.* (1990). Argument Structure. MA: MIT Press.
12. *Grimshaw J., Mester A.* (1988). Light verbs and θ-marking. Linguistic inquiry, vol.19 (2).

Семантическая структура слова и его взаимодействие с отрицанием[1]

Semantic Structure of Words and their interaction with negation

Апресян В. Ю. (valentina.apresjan@gmail.com)

Институт русского языка им. В. В. Виноградова, Москва

В работе рассматриваются факторы, влияющие на взаимодействие предикатов с отрицанием. На способность семантического компонента, находящегося в пресуппозиции, иногда попадать под отрицание, влияет его семантическая природа, а также его место в языковой картине мира. Утверждается, что экзистенциальные семантические компоненты могут переходить из пресуппозиции в ассерцию, а оценочные семантические компоненты — нет. Это связано с тем, что в языке оценки и мнения чаще имплицируются и составляют заданный фон высказывания, а утверждения о существовании и особенно не-существовании составляют ассерции высказываний, попадая в фокус внимания. Иногда очень разные свойства при взаимодействии с отрицанием обнаруживаются у близких синонимов (*повезло,* с одной стороны, и *посчастливилось*, с другой). Их семантические и синтаксические различия определяются логической структурой их семантических компонентов: у интерпретатива *повезло* экзистенциальный компонент находится в пресуппозиции (и легко выводится на поверхность, в ассерцию), а оценка — в ассерции, у предиката *посчастливиться* — оценка в пресуппозиции, а экзистенциальный компонент — в ассерции.

Семантические особенности предикатов, определяющие их способность подпадать под действие отрицания, определяют также возможные для них модели управления, за некоторыми из которых закреплен определенный тип взаимодействия с отрицанием.

Введение

Традиционно под пресуппозитивной частью значения слова понимается та частьего толкования, которая остается неизменной под отрицанием, а под ассерцией, или утверждением — та, которая меняется. Известны целые группы слов, содержащие пресуппозиции. Приведем несколько примеров. Так, в значение многих глаголов совершенного вида входит значение соответствующих глаголов несовершенного вида в качестве пресуппозиции; ср. известный пример *решать-решить (задачу)* [Апресян 1995:59–60], где значение НЕСОВ *решать* ‘Человек X обдумывал информацию, имеющую отношение к продукту мысли Y, с целью получить ответ на содержащийся в Y-е вопрос’ составляет пресуппозицию СОВ *решить*, в значение которого также добавляется ассертивный элемент ‘X получил ответ на этот вопрос’. Во фразе *Он не решил задачу* тот факт, что он ее решал, не подвергается отрицанию.

Другая крупная группа слов, в значении которых регулярно выделяются пресуппозиции, в числе прочих слоев смысла, — это глаголы со значением речевых актов (РА) типа *просить, требовать, обещать, сообщать, жаловаться* и т. д., у которых пресуппозиции «отражают мнения и желания субъекта РА, взаимные установки субъекта и адресата РА» [Гловинская 2010]. Ср., например, толкование глагола РА *вразумить* в той же работе, где пресуппозиция вводится придаточным предложением с ‘считая’ [Гловинская 2010]:

вразумить (*насилу вразумить кого-то*) ‘Человек А1, считая, что человек А2 неправильно ведет себя

[1] Работа выполнена при финансовой поддержке Программы фундаментальных исследований отделения историко-филологических наук РАН «Генезис и взаимодействие социальных, культурных и языковых общностей», а также гранта Президента Российской Федерации для государственной поддержки ведущих научных школ Российской Федерации НШ-4019.2010.6.

или не имеет правильного мнения о чем-то, объяснил ему это с помощью рассуждений А3; в результате А2 начал иметь правильное мнение или правильно себя вести'.

Хотя пресуппозиция, по логическому определению, не отрицается, существуют условия и контексты, в которых явно пресуппозитивная часть значения попадает в сферу действия отрицания. Это явление рассматривается, в частности, в работах Е. В. Падучевой. В работе [Падучева 2005], на собственных примерах автора и примерах, привлеченных из других работ, формулируются условия, когда это может происходить: во-первых, распространению отрицания на пресуппозицию способствует контекст снятой утвердительности (будущее время, повелительное наклонение и др.), во-вторых, «граница между пресуппозицией и ассерцией менее отчетливая и отрицание легко ее преодолевает» в случае наречных сочетаний с адвербиалами типа *вовремя (прийти), второпях (оставить ключи), резко (затормозить), дешево (продать)* и пр.

В работе [Кустова 1996] рассматривается влияние коммуникативных факторов, способствующих распространению отрицания на семантические элементы пресуппозитивного характера. В работе рассматриваются, в частности, глаголы эмоциональной каузации типа *огорчить, обидеть, расстроить, испугать, обрадовать*, имеющие в качестве пресуппозиции указание на совершение некоторого поступка, а в качестве ассерции — указание на эмоциональную реакцию. Автор демонстрирует возможность двух интерпретаций для фраз типа *Петя не огорчил маму двойкой* — 'Он не получил двойку, и мама не огорчилась'; 'Он получил двойку, но мама не огорчилась'. Возможность двух интерпретаций отрицательных предложений связывается с наличием двух возможных коммуникативных структур для исходного предложения с личной конструкцией 'X совершил поступок P, который каузировал Q' (Петя получил двойку, и это огорчило маму) и 'Поступок P, который совершил X, каузировал Q' (Получение двойки Петей огорчило маму). Отрицание первой структуры дает интерпретацию с широкой сферой действия, отрицание второй структуры — с узкой. При этом пресуппозицию 'X совершил поступок' автор квалифицирует не как лексическую, т. е. закрепленную за семантикой слова, а как текстовую.

Еще одним примером возможности распространения отрицания на пресуппозитивную часть толкования являются интерпретативы — предикаты, которые обозначают не конкретное действие, деятельность или состояние, а их интерпретацию говорящим; ср., например, этическую интерпретацию (*помогать, предавать, издеваться*), логическую интерпретацию (*ошибаться, просчитаться*) и т. п. [Апресян 2006:150]. Семантическое устройство интерпретативов таково — их пресуппозицию образует указание на конкретное действие, а ассерцию — квалификация этого действия говорящим [Апресян 2006:151]; ср. *Он не ошибается, думая, что его собираются уволить* (отрицается только наличие ошибки, но не наличие определенного ментального состояния). Ср. семантическую структуру интерпретатива *совершать грех* по работе [Апресян 2006:149]:

X совершает грех, делая P = 'X делает [пресуппозиция]; говорящий считает, что P относится к классу действий, которые Бог запрещает людям совершать [ассерция]'.

При этом автор отмечает, что прототипическим для интерпретативов является употребление в форме НЕСОВ НАСТ, и что в контекстах СОВ ПРОШ указание на действие переходит в ассерцию и, соответственно, может попадать под отрицание.

Отметим, что переход указания на действие в ассерцию в форме СОВ ПРОШ не является обязательным: фразы типа *Он не согрешил, изменив ей* могут иметь по две интерпретации — 'Он не изменил ей' [указание на поступок в ассерции] и 'Говорящий не считает, что измена ей — это грех' [указание на поступок в пресуппозиции], причем вторая интерпретация является предпочтительной. Однако в отсутствие деепричастного оборота переход пресуппозиции в ассертивную часть облегчается: фразу *Он ее не предал* естественнее интерпретировать как 'Он не совершил никакого поступка, который можно квалифицировать как предательство', чем как 'Тот поступок, который он совершил, нельзя квалифицировать как предательство'.

В данной работе на примере двух близких по значению предикатов — *повезти* и *посчастливиться* — рассматриваются некоторые факторы, влияющие на возможность перехода пресуппозиции в ассерцию, а именно, семантическая природа пресуппозитивного компонента. Утверждается, что экзистенциальные компоненты могут переходить в ассерцию, в то время как для оценочных компонентов это невозможно. Это объясняет возможность отрицания пресуппозитивных элементов у эмоциональных каузативов и у интерпретативов, где в пресуппозиции находится указание на совершение некоторого поступка, и невозможность их отрицания у речевых актов, где в пресуппозиции входят разные виды мнений и оценок. Кроме того, на возможность компонента подвергаться отрицанию влияет его место в языковой картине мира: если семантический компонент указывает на какие-то идеи, укорененные в языковой картине мира, это резко снижает его способность подвергаться отрицанию.

Повезти Vs. *посчастливиться*

Глагол *повезти (везти)* имеет интерпретативное значение, в котором он употребляется в безличной форме: *Ему всегда везет*; *Ему повезло, что*

он с ней встретился; *Это ему повезло*; *Ему с ней повезло*; *Ему повезло работать с этим режиссером*; *Ему не повезло с работой*; *Ему повезло: он работает с таким режиссером!*.

Он имеет следующую семантическую структуру:

X-у повезло с Y-ом <повезло, что Y; повезло Y-ть; повезло: Y; Y. — Это X-у повезло> = 'С X-ом произошло событие Y [пресуппозиция]; говорящий считает, что Y произошел не в результате действий X-а, а в результате стечения обстоятельств [пресуппозиция]; говорящий считает, что Y — это хорошее событие'[2].

Поясним предложенное для *повезти* толкование. Оно, как и у других интерпретативов, содержит в пресуппозиции указание на событие, а в ассерции — определенную квалификацию этого события. Ср. фразу *Им здорово не повезло, что такое громкое дело повесили на них* (В. Громов), где отрицается квалификация события как хорошего, но не само его наличие.

При этом у *повезти* есть и некоторые особенности по сравнению с другими интерпретативами. Часть квалификативной части у него входит в пресуппозицию, а именно квалификация события как результата стечения обстоятельств, а не результата собственных действий X-а. Во фразе *Им здорово не повезло, что такое громкое дело повесили на них* не отрицается ни сам факт события, ни его случайность, только его оценка как хорошего.

Следует при этом отметить любопытный статус двух пресуппозиций слова *повезло*. С одной стороны, они отвечают критерию пресуппозитивности, то есть, по крайней мере, в большинстве контекстов не подвергаются отрицанию. С другой стороны, они не являются абсолютно недоступными для семантического взаимодействия с другими элементами высказывания, в отличие, скажем, от пресуппозиций речевых актов. Рассмотрим пресуппозицию глагола *вразумить*, упоминавшегося выше: 'Человек A1, считая, что человек A2 неправильно ведет себя'. Она не может попадать в семантическую сферу действия других единиц высказывания в силу того, что можно назвать ее «глубоким семантическим залеганием»: ср. неправильность **Он ее очень вразумляет, *Он ее абсолютно вразумляет*. В этих фразах адвербиалы по смыслу могли бы относиться только к пресуппозитивной части 'человек A2 **очень / абсолютно** неправильно ведет себя', однако в силу «глубокого залегания», «запрятанности» пресуппозиции, они не могут вступить с ней в семантическое взаимодействие. Ассертивный элемент глагола *вразумить*, 'человек X объяснил' также не может градуироваться адвербиалами *очень* и *абсолютно*, что приводит к семантической аномалии.

Что касается пресуппозиций глагола *повезло*, то они обе не слишком «глубоко залегают» — и поэтому, как и ассерция, доступны для семантического взаимодействия. В частности, они могут определяться адвербиалами с фиксированной семантической сферой действия — теми, которые «выбирают» компонент, с которым они могут семантически сочетаться [Филипенко 1998].

Ср., например, наречия со значением частотности и вероятности, требующими наличия компонента 'событие' в своей семантической сфере действия; для них доступна пресуппозиция 'произошло событие': *Ему **часто** везет, Ему **опять** повезло, **Авось**, повезет* ['**часто** происходит событие' '**опять** произошло событие', '**авось**, произойдет событие']. Адвербиалы со значением идентификации типа *просто, всего лишь* «выбирают пресуппозицию» 'в результате стечения обстоятельств': *Ему просто <всего лишь> повезло* ['**просто** в результате стечения обстоятельств', '**всего лишь** в результате стечения обстоятельств']. При этом степенные адвербиалы определяют ассерцию, в которой находится градуируемый компонент оценки: *Ему крупно <очень> повезло* ['Y– это **очень** хорошее событие'].

Интересно, что при этом статус у второй пресуппозиции 'в результате стечения обстоятельств' немного иной, нежели у первой пресуппозиции 'произошло событие', потому что на вторую пресуппозицию могут воздействовать и адвербиалы с плавающей сферой действия, т. е. не имеющие строгих семантических ограничений на сочетаемость с предикатами. Так, во фразе *Ему явно повезло* с интерпретацией 'произошло **явно** случайное и **явно** хорошее событие' в сферу действия адвербиала *явно* попадает не только ассертивный элемент 'хорошее событие', но и пресуппозитивный элемент 'в результате стечения обстоятельств'; следовательно, эта пресуппозиция «ближе к поверхности», чем пресуппозиция 'произошло событие'.

Более того, многие адвербиалы с фиксированной сферой действия, которые по смыслу могли бы сочетаться с первой пресуппозицией — 'произошло событие', дают в сочетании с глаголом *повезло* аномальные фразы. Так, фразы **Ему постепенно повезло, *Ему повезло за последние годы, *Ему быстро <стремительно> повезло* невозможны, при том, что фразы *Это произошло постепенно, Это произошло за последние годы, Это произошло быстро <стремительно>* абсолютно грамматичны, и, следовательно, смысл 'произойти', содержащийся в пресуппозиции *повезти*, в принципе сочетается

---

[2] Предлагаемое нами толкование очень близко толкованию, предложенному в работе [Зализняк, Левонтина 1996], где содержится много интересных наблюдений над словами триады *довелось — повезло — посчастливилось*; отличие состоит в том, что мы постулируем две обязательных валентности для слова *повезло* — субъекта и события (*повезти* может только *кому-то*, это предикат содержит именно субъективную оценку), однако логические структуры предлагаемых толкований практически идентичны; ср. толкование, предложенное Зализняк и Левонтиной:

[през.] P произошло
P — результат случайности
[асс.] (для X-а) хорошо, что P.

с подобными наречиями. Аномалия при сочетании с *повезти* возникает из-за того, что адвербиалы *постепенно, быстро, в последние годы* и пр. привлекают внимание не к самому факту события, а к тому, как оно протекало, а это, очевидно, представляет из себя недопустимо высокую степень фокусирования внимания на пресуппозиции.

Однако если при взаимодействии с адвербиалами пресуппозиция 'событие случайно' проявляет себя как более «близкая к поверхности», а пресуппозиция 'произошло событие' как более «глубокая», то с отрицанием эти пресуппозиции взаимодействуют прямо противоположным образом. Как и другие интерпретативы, *повезти* допускает переход пресуппозиции 'произошло событие' в ассерцию под отрицанием: *Ему не повезло родиться богатым* 'Он не родился богатым'[3]. Однако вторая пресуппозиция, а именно, 'событие случайно' не может отменяться под отрицанием: когда мы говорим *Ему не повезло родиться богатым, Ему не повезло стать победителем*, мы отрицаем, что нечто произошло, но не отрицаем случайный характер этого хода событий.

Таким образом, с одной стороны, элемент 'Y произошел не в результате действий X-а, а в результате стечения обстоятельств' доступен для воздействия адвербиалов, с другой стороны, он не поддается отрицанию. Это может быть связано с особенностями русской языковой картины мира, в которой сильно представление о том, что существует некоторая внешняя по отношению к человеку сила, не контролируемая им и определяющая ход событий; ср. об этом [Wierzbicka 1990, Булыгина, Шмелев 1997, Зализняк и Левонтина 1996, Шмелев 2005, Апресян 2006, В. Апресян 2008]. Пресуппозитивный, неотрицаемый характер элемента 'Y произошел в результате стечения обстоятельств' соответствует укорененности идеи 'судьбы' в русской языковой картине мира.

Что касается отрицания пресуппозиции 'произошло событие Y', то здесь действуют как семантические, так и синтаксические факторы.

Как видно из сентенциального входа и примеров, у *повезти* достаточно разнообразные возможности реализации валентности P — *что*-предложение, бессоюзие, инфинитив, предложно-именная группа *с* ТВОР, анафора со словами типа *это*[4]. При этом не все из них позволяют переход пресуппозиции 'произошло событие Y' в ассерцию под отрицанием. Только две модели управления являются однозначными, а именно, предложение, вводимое союзом *что* и бессоюзие, которые всегда интерпретируется фактивно — пресуппозиция не попадает в сферу действия отрицания: фразы *Ему не повезло, что он родился богатым* и *Ему не повезло: он родился богатым* могут значить только 'Он родился богатым, и это плохо'.

Прочие модели управления ведут себя под отрицанием следующим образом. Инфинитив предпочтительно интерпретируется с широкой сферой действия, т. е. с переходом пресуппозиции в ассерцию: *Не повезло стать директором* 'Не стал директором', *Что делать, если не повезло родиться красивой* 'Если не родилась красивой', *Те, кому не повезло работать у хороших работодателей, вынуждены тратить на повышение квалификации собственные средства* 'Не работают у хороших работодателей'.

В узусе имеется достаточно большое количество примеров с инфинитивом, где пресуппозиция не попадает в сферу действия отрицания; ср. *Мне не повезло работать в чисто женском коллективе* ['Я работаю в чисто женском коллективе, и это плохо'], *Ему не повезло родиться в семье алкоголика* ['Он родился в семье алкоголика, и это плохо'], однако многими образованными носителями языка эти примеры воспринимаются как находящиеся за гранью литературной нормы. Таким образом, для инфинитива под отрицанием в целом предпочтителен переход пресуппозиции в ассерцию, хотя для какой-то части носителей языка примера типа *Ему не повезло родиться в Америке* являются амбивалентными, особенно в контекстах с нефиксированной оценкой типа *Ему повезло — или не повезло — родиться в Америке*[5].

Конструкция *с* ТВОР по умолчанию фактивна, однако в специально сконструированных контекстах допускает отрицание пресуппозиции: ср. *Ему не повезло с женой* 'У него есть жена, и эта жена плохая', но *Ему не повезло с богатой женой* 'У него нет богатой жены, и это плохо'.

Конструкция *это <в этом>* одинаково свободно допускает обе интерпретации: *–Меня поселили в номере с Ивановой. — Что ж, это <в этом> тебе*

---

[3] Некоторые носители языка допускают две интерпретации для этой фразы — с широкой сферой действия отрицания, фокусом внимания и фразовым ударением на *повезло* (*Ему ↓не повезло родиться богатым* 'Он не родился богатым'), а также узкой сферой действия отрицания, фокусом внимания и фразовым ударением на *родиться богатым* (*Ему ↓не повезло родиться богатым* 'Он родился богатым, и это плохо').

[4] В том, что касается синтаксических свойств предиката *повезло*, наша трактовка расходится с предлагаемой в работе [Зализняк, Левонтина 1996], где утверждается, что «у *повезло* валентность содержания является факультативной: она никогда не заполняется зависимым инфинитивом, редко — придаточным с союзом *что*, но может и вообще оставаться незаполненной». В НКРЯ встретилось 96 случаев заполнения валентности содержания у *повезло* инфинитивом и 235 — *что*-предложением, при большом разнообразии прочих синтаксических средств заполнения — бессоюзия, анафоры, группы *с* ТВОР. Таким образом, нельзя утверждать, что для *повезло* предпочтительны эллиптические конструкции, в невыраженной валентностью содержания.

[5] Примеры выше в принципе также оценочно неоднозначны, т. к. и у того, чтобы работать директором, и у того, чтобы быть красивой, нетрудно найти отрицательные стороны — и то, и другое может быть хлопотным и опасным, отчего становится возможной двоякая интерпретация.

не повезло ‘Поселили с Ивановой, и это плохо’; *Меня не поселили с Петровой! — Что ж, это <в этом> тебе не повезло* ‘Не поселили с Петровой, и это плохо’.

Интересный контраст с глаголом *повезти* составляет его близкий синоним *посчастливиться*, на котором видно, как особенности семантической структуры глагола влияют на его взаимодействие с отрицанием.

На первый взгляд, *повезти* и *посчастливиться* — близкие синонимы; ср. *Ему повезло иметь такого учителя как Станиславский*; *Ему посчастливилось иметь такого учителя как Станиславский*. И в случае *повезти*, и в случае *посчастливиться* событие P оценивается хорошее и как произошедшее не в результате действий X-а.

Однако прагматический фон, семантика, аспектуальные и синтаксические свойства у этих двух глаголов очень разные.

Во-первых, *посчастливиться*, в отличие от *повезти*, обычно описывает обычно однократное (или, во всяком случае, достаточно редкое), и при этом очень важное событие: *Надеюсь, что в этой небольшой главе мне хоть в малой степени удалось рассказать об этих удивительных людях, личностях, с которыми мне посчастливилось встретиться в своей жизни* (И. Архипова); *Главное — мне посчастливилось общаться и дружить с воистину легендарными людьми* (С. Спивакова). *Посчастливиться* — это то, что происходит далеко не с каждым человеком, это выход за пределы некоторой нормы, и часто — это некоторое глобальное событие.

Повезти — это гораздо более локальное событие, поэтому *повезти* может *в чем-то*. Сочетание *посчастливиться в чем-то* малоупотребительно, именно потому, что *посчастливиться* предполагает гораздо более глубокое и глобальное воздействие события на жизнь человека.

Кроме того, *повезти* существует на фоне своего отрицания — *не повезти*[6]. Этот концепт содержит представление о том, что в каждый момент, в том числе, когда происходит что-то важное, человеку может *повезти* или *не повезти*. *Везучим* людям чаще *везет*, *невезучим* — чаще *не везет*, однако в целом среднестатистическому человеку иногда *везет*, а иногда — *не везет*. Более того, часто бывает, что человеку в *чем-то везет* (например, *в любви*), а в чем-то — *не везет* (например, *в картах*). *Посчастливиться* же — это редкая улыбка судьбы.

Наконец, *посчастливиться* — это некоторое положительное событие: если глаголом *повезти* можно обозначить отсутствие возможного плохого события, то глагол *посчастливиться* в такой ситуации неуместен. Ср. *Ему еще повезло, что левую руку сломал, а не правую*; *Ему еще повезло, что дом не сгорел*, при крайней прагматической странности ??*Ему еще посчастливилось сломать левую руку, а не правую*; ?*Ему еще посчастливилось, что дом не сгорел*. Если вероятность какого-то очень плохого события была очень высока, так что избавление от него можно считать редкой удачей, то глагол *посчастливиться* становится более уместным: *Ему посчастливилось не погибнуть в этом страшном бою <выйти живым из окружения>*; *Там много положили их, безоружных партизан, редко кому посчастливилось вынести целой собственную голову* (В. Быков).

Таким образом, *посчастливиться* — редкое, экстраординарное, исключительно хорошее и глобально влияющее на жизнь человека событие[7]. Как мы видим, *посчастливиться* содержит больше оценочных компонентов, их удельный вес в толковании больше. Как видно из предлагаемого толкования, у *повезти* и *посчастливиться* также отчасти различается и их статус в логической структуре значения.

Мы предлагаем следующее толкование для глагола *посчастливиться*:

X-у посчастливилось, (что) Y = ‘говорящий считает, что Y произошел не в результате действий X-а, а в результате вмешательства внешних по отношению к человеку сил [пресуппозиция]; говорящий считает, что такие события, как Y происходят очень редко [пресуппозиция]; говорящий считает, что Y — это очень хорошее событие [пресуппозиция]; с X-ом произошло событие Y [ассерция]’.

---

[6] Интересно, что, например, в английском языке *повезти* и *не повезти* — это разные лексемы *to have good luck* и *to have bad luck*, общим у которых является смысл ‘случайное событие’, а оценка «вынесена» в отдельное слово — часть фраземе. Впрочем, само по себе слово *luck* содержит слабый смысл ‘хорошее событие’, который исчезает при добавлении слов с отрицательной оценкой — *bad*, *ill*. Интересно, что в тех случаях, когда *не повезти* имеет широкую сферу действия, и отрицается не только оценка события как хорошего, но и сам факт события как такового, в английском используются другие единицы — *not get lucky, not have luck*. Таким образом, в английском нет той амбивалентности, которая возможна в русском: в качестве коррелята к фразам с русской узкой сферой действия, где отрицается положительная оценка, используется одна языковая единица (*Ему не повезло, что он сломал ногу = He had the* ***bad luck*** *to break his leg*), а качестве коррелята к фразам с русской широкой сферой действия, где отрицается само событие, используется другая языковая единица (*Ей не повезло родиться красавицей = She* ***didn’t have the luck*** *to be born a beauty*).

[7] *Для посчастливиться* прототипическим является именно событие, а не действие, как и для всех глаголов этого ряда; ср. *Мне посчастливилось жить в одно время с этим великим человеком*, но не **Мне посчастливилось написать отличную книгу.* В этом наше понимание *посчастливиться* расходиться с предложенным в работе [Зализняк, Левонтина 1996], где *посчастливиться* толкуется через действие (хотя мы принимаем предложенную в работе логическую структуру смысла этого глагола, где оценка находится к пресуппозиции, а событие — в ассерции).

X-у посчастливилось сделать P — [през.] вероятность P была мала
P хорошо
P произошло как бы само собой
[асс.] X сделал P

Как мы видим, у *посчастливиться* указание на событие содержится в ассерции, а указание на оценку события как хорошего — в пресуппозиции. Т.е., когда мы говорим *не посчастливилось Y*, то мы отрицаем сам факт события Y, а когда мы говорим *не повезло Y*, то мы можем отрицать как сам факт события Y, так и его оценку как хорошего. Указание на случайность события у обоих глаголов содержится в пресуппозиции; кроме того, у *посчастливиться* есть дополнительная пресуппозиция — указание на редкость, на из ряда вон выходящий характер события. При этом все пресуппозиции глагола *посчастливиться* неотрицаемы.

В силу своих семантических и прагматических особенностей *посчастливиться*

а) не имеет, в отличие от *повезти*, формы НЕСОВ, т. к. не может обозначать многократных и узуальных событий (в силу своей моментальности *посчастливиться*, как и *повезти*, не может также иметь и актуально-длительных и процессных значений): **Ему иногда счастливится*; **Ему обычно счастливится*;

б) не управляет, в отличие от *повезти*, предложно-именной группой *с* ТВОР, указывающей на тот аспект, в котором проявилось везение: всего 6 вхождений *посчастливилось с чем-л.* в НКРЯ на сотни вхождений *повезло с чем-л.*;

в) не управляет, в отличие от *повезти*, анафорическим местоимением *это*: в НКРЯ встречаются единичные фразы типа *Это тебе посчастливилось*, однако в принципе это интерпретационное употребление противоречит прагматике *посчастливиться*, которое, как правило, указывает на столь важное, хорошее и экстраординарное событие, что не может быть никаких сомнений в том, как его надо интерпретировать;

г) гораздо меньше, чем *повезти*, сочетается с отрицанием (*не посчастливилось* составляет всего около 9 процентов от всех употреблений этого глагола; *не повезло* — примерно 25 процентов). Если и присутствие, и отсутствие *везения* находится в пределах нормы или отклоняется от нее лишь незначительно, то *посчастливиться* — это нечто, сильно выходящее за пределы среднестатистической нормы. Нормой является отсутствие этого явления, которое поэтому и не требует никаких специальных средств выражения.

В тех случаях, когда *посчастливиться* все-таки сочетается с отрицанием, этот глагол также ведет себя иначе, нежели *повезти* (см. также выше).

Во-первых, под отрицанием для *посчастливиться* невозможна фактивная интерпретация, отчего этот глагол под отрицанием не может управлять фактивной *что*-пропозицией: *Ему не посчастливилось вновь с ней встретиться* ‘Он с ней не встретился’, но не **Ему не посчастливилось, что он с ней не встретился*. Это дополнительное подтверждение того, что у *посчастливиться* указание на событие находится не в пресуппозиции, как у *повезло*, а в ассерции, поскольку нет условий, в которых оно могло бы сохраняться под отрицанием.

Во-вторых, у *посчастливиться* под отрицанием не отменяется оценка события как хорошего.

Мы видим, таким образом, что экзистенциальные пресуппозитивные компоненты, такие как ‘произошло событие Y’ у глагола *повезти*, ‘X совершил поступок Y’ у эмоциональных каузативов (*Петя огорчил маму двойкой*) и у интерпретативов типа (*Он согрешил, изменив ей*) при определенных условиях могут переходить из пресуппозиции в ассерцию: ср. *Ему не повезло занять первое место* (Он его не занял), *Петя ни разу не огорчил маму двойкой* (Не получил двойку), *Он ни разу не согрешил, изменив ей* (Он не изменил ей). В то же время, оценочные пресуппозитивные элементы отрицанию не подвергаются: в частности, никогда не отрицается оценка события как хорошего, редкого и случайного у *посчастливиться*, оценка события как случайного у *повезти*, оценки статусов, ситуаций и адресатов у речевых актов. Это может быть связано с тем, что в языке оценки и мнения чаще имплицируются и составляют заданный фон высказывания, а как утверждения о существовании и особенно о не-существовании составляют ассерции высказываний, попадая в фокус внимания.

Литература

1. *Апресян Ю. Д.* Избранные труды. Т. II. Интегральное описание языка и системная лексикография. // М.: Языки русской культуры. 1995. С. 348–386.
2. *Апресян Ю. Д.* Основания системной лексикографии // Языковая картина мира и системная лексикография. Отв. ред. Ю. Д. Апресян. М.: «Языки славянских культур», 2006. С. 145–160.
3. *Апресян В. Ю.* О судьбе и не-судьбе // Динамические модели: слово, предложение, текст. Сб. статей в честь Е. В. Падучевой. М.: Языки славянских культур, 2008. С. 7–19.
4. *Булыгина Т. В., Шмелев А. Д.* Языковая концептуализация мира (на материале русской грамматики) // М.: Языки русской культуры, 1997. С. 200–207.
5. *Гловинская М. Я.* Лексикографические типы: речевые акты и травмы // Проспект Активного словаря русского языка под общим руководством академика Ю. Д. Апресяна (в печати).
6. *Зализняк Анна А., Левонтина И. Б.* Отражение «национального характера» в лексике русского языка. Размышления по поводу книги: Anna Wierzbicka. Semantics, Culture, and Cognition. Universal Human Concepts in Culture-Specific Configurations. N.Y., Oxford, Oxford University Press, 1992 //Russian Linguistics, 1996, vol. XX.
7. *Зализняк Анна А., Левонтина И. Б., Шмелев А. Д.* Ключевые идеи русской языковой картины мира. // М.: Языки славянской культуры, 2005.
8. *Кустова Г. И.* О коммуникативной структуре предложений с событийным каузатором // Московский лингвистический журнал. Т. 2. М.: РГГУ, 1996. С. 240–261.
9. *Падучева Е. В.* Эффекты снятой утвердительности: глобальное отрицание // Русский язык в научном освещении. 10 (2). М.: «Языки славянских культур», 2005. С. 17–42.
10. *Филиппенко М. В.* Об адвербиалах с плавающей и фиксированной сферой действия (к вопросу об актантах и не-актантах предиката) // семиотика и информатика. Вып. 36. М.: «Языки русской культуры», «Русские словари», 1998. с. 120–140.
11. *Wierzbicka A.* Dusa ‘soul’, toska ‘yearning’, sud’ba ‘fate’: three key concepts in Russian language and Russian culture // Metody formalne v opisie jezykow slowianskix, ed. Zygmunt Saloni. Dzial Wydawnitctw Filii UW w Bialymstoku, 1990.

Семантика фразеологизмов: иерархия или сеть?

Semantics of idioms: hierarchy or semantic net?

Баранов А. Н. (baranov_anatoly@hotmail.com),
Добровольский Д. О. (dm-dbrv@yandex.ru)

Институт русского языка РАН

Рассматриваются два основных подхода к идеографическому описанию фразеологии: индуктивный и дедуктивный. Приводятся аргументы в пользу индуктивного подхода. В качестве иллюстративного материала используется «Словарь-тезаурус современной русской идиоматики»

1. Моделирование семантических отношений в лексике

Идеографическое описание лексики допускает два основных подхода к анализу языковых данных: индуктивный и дедуктивный. Дедуктивный подход основывается на представлении, что существует некая априорная классификация сущего. В этом случае искусство лексикографа состоит в том, чтобы приложить эту схему к языковому материалу. Источники такой классификации формально должны были находиться за пределами лингвистики — в сфере философии и логики. Логическим по сути является, в частности, «дерево Порфирия». В систематике — науке, оформившейся в конце XVI века, — возникло представление о «естественной системе» классификации, которая оказалась теоретическим обоснованием идеи единственности правильной классификации.

Лингвистические тезаурусы, разрабатывавшиеся с середины XIX века, ориентировались на то, что такая схема должна быть. Эта идея отнюдь не умерла и сегодня: большинство современных тезаурусов исходит из существования такой иерархической системы отношений, которая может быть наложена на язык «сверху». Отметим, что стратегия навязывания языковому материалу жесткой логической схемы противоречит самому духу современной лингвистики. Языковые феномены настолько сложны, что описание их с помощью заранее заданных фиксированных противопоставлений существенно огрубляет наши знания об исследуемом феномене.

Дедуктивному методу построения тезауруса может быть противопоставлен метод индуктивный, основанный на семантическом анализе конкретного материала. Понятно, что в этом случае не обязательно возникает последовательная иерархия понятий по принципу «от более частного к более общему», т. е. некая логически непротиворечивая древовидная структура категорий. Как показывает практика нашей работы над «Словарем-тезаурусом современной русской идиоматики» (далее — Тезаурус), в результате получается множество «деревьев» относительно небольшой глубины. Более того, эти «деревья» связаны между собой, причем не отношениями «род — вид», а другими семантическими отношениями, которые не менее важны. Такая структура с формальной точки зрения больше напоминает семантическую сеть. Это отражает естественное устройство семантики фразеологии, которая в общем случае организована не по иерархическому, а по сетевому принципу[1]. Связи между таксонами отражаются в виде парадигматических отсылок[2]. Отсюда следует, что нет мотивированного порядка расположения таксонов относительно друг друга, хотя чисто технические способы упорядочивания таксонов здесь необходимы.

Отметим, что сетевая структура построения идеографического словаря более оправдана не только по отношению к фразеологии, но и ко всей лексике в целом.

---

[1] О сетевых моделях семантической организации лексики см., например, [Quillian 1968].

[2] Интересно, что аналогичные идеи о структуре словарей-тезаурусов высказывались разработчиками информационно-поисковых тезаурусов. Ср., например, [Варга 1970].

2. Семантические отношения в идеографическом описании фразеологии

В общем случае семантическая сеть формируется узлами, которые соединяются между собой различными семантическими отношениями. В отличие от дерева, семантическая сеть не ориентирована, не имеет иерархической структуры. В зависимости от типа семантических отношений между узлами могут существовать отношения наследования семантических свойств (для фрагментов деревьев), каузальные связи, отношения включения и пр. С формальной точки зрения Тезаурус представляет собой совокупность таксонов, упорядоченных по достаточно произвольным характеристикам. Как известно, отдельные фрагменты лексической системы могут быть упорядочены в виде дерева — например, термины родства, названия цветов, овощей, фруктов, животных и пр. «Древесная» структура может быть обнаружена и во многих таксонах Тезауруса (как уже говорилось в предыдущем разделе). По большей части глубина этих таксонов невелика и не превышает двух-трех уровней, однако в отдельных случаях может достигать и пяти уровней (ср. поля ВРЕМЕНИ и КОЛИЧЕСТВА). Древовидные таксоны содержат корневую часть, в которую включены идиомы с очень общей семантикой, а также идиомы, для которых не нашлось особого подтаксона в рамках данного древоводного таксона.

Традиционные структуры тезаурусов в том виде, в котором они представлены в словарях Роже [Roget's Thesaurus 1987], Дорнзайфа [Dornseiff 1959], Халлига и Вартбурга [Hallig, Wartburg 1963], — это компромисс, поскольку построение функции отображения сети (т. е. семантической структуры языковых форм) в дерево приводит к неизбежным искажениям. Следовательно, в тезаурусах необходимы «горизонтальные» отсылки (ср. фиксацию таких связей во вполне традиционном тезаурусе Роже). Как уже было сказано выше, в Тезаурусе связи между таксонами выражаются с помощью парадигматических отсылок. Еще одним способом отражения сетевых связей в идиоматике является разделение таксона на центральную и периферийную часть (см. ниже раздел 3).

В информационно-поисковых тезаурусах, которые образуют основу информационно-поисковых языков, обычно выделяются три основных вида парадигматических отношений: отношения «род-вид», антонимические и ассоциативные отношения. Поскольку выделение ассоциативных отношений трудно формализовать, а отсутствие ясных критериев может привести к абсолютному произволу при создании системы парадигматических отсылок, то в основу выделения таксонов были положены именно родо-видовые отношения и их модификации. Заметим, что антонимические отношения и часть родо-видовых отношений фиксируются в самой структуре таксонов Тезауруса. Например, отношение антонимии представлено в названиях таксонов ХОРОШО — ПЛОХО, БЕДНОСТЬ — БОГАТСТВО, ДВИЖЕНИЕ — ОСТАНОВКА, ВАЖНОСТЬ — НЕВАЖНОСТЬ.

В Тезурусе отражались следующие типы парадигматических отношений. Отношение **включения**, предполагающее, что идиомы с более «частным» значением помещаются в соответствующие им таксоны, а от таксонов с более общим значением к ним делаются отсылки. Например, идиомы *до седых волос, старая гвардия, старая дева, мышиный жеребчик, старой закваски* включены в таксон СТАРОСТЬ, а из таксона с более общим смыслом ДАВНО сделана отсылка к таксону СТАРОСТЬ. Большое количество отсылок такого типа к другим таксонам содержат поля ИНТЕНСИФИКАТОРЫ ОБЩЕГО ХАРАКТЕРА, ХОРОШО, ПЛОХО. Это связано с тем, что такие смыслы, как 'очень, сильно', 'хорошо' и 'плохо' представлены в семантике большого количества идиом, распределенных по самым разным таксонам. Понятно, что отсылки, за которыми стоит отношение включения, однонаправлены, поскольку само отношение включения асимметрично.

Классические родо-видовые отношения отличаются от отношения включения только тем, что категория X является родовой по отношению к категориям Y, …, Z. В качестве примера родо-видовых отношений приведем таксоны НАКАЗАНИЕ… и КАЗНЬ. Поскольку КАЗНЬ (ср. *высшая мера [наказания], пустить… в расход, поставить к стенке*) является видом НАКАЗАНИЯ, в таксоне НАКАЗАНИЕ… есть отсылка к КАЗНИ, но не наоборот. Такая направленность семантических отношений (от более общего — к более частному) объясняется тем, что читатель тезауруса, который интересуется идиомами более частной семантики вряд ли будет интересоваться более общими смыслами, связанными с данным таксоном. Например, читатель, ищущий идиомы поля ХОРОШАЯ ОДЕЖДА; КРАСИВАЯ, КАЧЕСТВЕННАЯ И НОВАЯ ОДЕЖДА; ХОРОШО ОДЕТЫЙ вряд ли будет интересоваться идиомами из более общего поля ХОРОШО. В то же время тот, кто интересуется идиомами с семантикой хорошего вполне может заинтересоваться и идиомами, описывающими хорошую одежду. Данное решение, конечно, ориентировано на о б ы ч н о г о читателя о б ы ч н о г о словаря. Электронный тезаурус практически (с точки зрения лингвиста) не ограничен в количестве возможных отображаемых связей, но это уже другой лексикографический продукт с другими задачами и даже с иной структурной организацией.

В некоторых случаях отражалась не безусловная логико-семантическая связь между таксонами, а скорее вероятностная. Здесь можно говорить об **отношении пересечения по экстенсионалу.** Например, некоторые идиомы таксона СТАРОЕ типа

сдать в архив, видавший виды, пропахнуть нафталином, в обед сто лет могут использоваться по отношению к старым людям (хотя и не обязательно). Иными словами, часть экстенсионала таксона СТАРОЕ пересекается с частью экстенсионала таксона СТАРОСТЬ. И наоборот — некоторые идиомы таксона СТАРОСТЬ (*старая гвардия, старой закваски*) в некоторых контекстах обращают внимание не на старость человека, а на его воззрения, нормы, ценности, которые интерпретируются как старые. В таких случаях делались симметричные отсылки.

В выбранном нами способе организации Тезауруса сетевые отношения эксплицируются не только парадигматическими отсылками, но и дублированием идиомы в разных таксонах. Это может указывать как на то, что идиома употребляется в этих таксонах в разных значениях, так и на то, что в одном и том же значении в ее плане содержания представлены разные семы. Например, речевая формула *снова(-) здорово* содержит в своей семантике компоненты ‘опять, повторно’ и ‘проблемы, трудности’. Ср. следующее толкование ее значения: ‘выражение недоумения и недовольства неожиданным появлением проблемы, которая ранее, предположительно, была решена или потеряла свое значение с течением времени, *в форме повторного приветствия этой проблемы как вновь пришедшего человека*’.[3] Тем самым эта речевая формула попадает в таксоны СНОВА, ОПЯТЬ и ТРУДНОСТИ…

3. Отражение различных слоев семантики в идеографическом представлении фразеологии

Как известно, план содержания языковых выражений неоднороден с точки зрения значимости его отдельных компонентов [Апресян 1995; Падучева 1977; Wierzbicka 1996]. Это в полной мере относится и к идиомам. Например, идиома *поживём — увидим* содержит в своем значении компоненты, которые соотносятся и с семантикой ‘неопределенности’ и с семантикой ‘будущего’, причем значимость этих смысловых блоков различна. Судя по контекстам типа (1) *Адвокат послал дело Васильева в Москву. Кто и как будет читать его там,* ***поживем — увидим***. [Корпус Публ.] эта идиома употребляется в ситуации, когда, говорящий не знает или не уверен, что будет происходить в будущем.

В плане содержания этой идиомы выделяется собственно пропозициональный компонент и его интерпретация. В данном случае коммуникативно более важным является именно пропозициональный компонент, а интерпретация вторична. Дело в том, что интерпретация описываемого события в значении данной идиомы включена во внутреннюю форму, т. е. в составляющую плана содержания идиомы, на основе которой формируется ее актуальное значение. Можно предложить следующее толкование этой идиомы: ‘Ответная реакция на вопрос или высказывание собеседника (в том числе на вопрос, поставленный самому себе), выражающая нежелание говорящего определенно отвечать на поставленный вопрос или выражать свое отношение к высказыванию собеседника из-за незнания ответа или из-за сомнений в правильности высказывания, а также по другим причинам (например, из-за нежелания обнаруживать свое мнение или информированность) *в форме утверждения о том, что последующая жизнь позволит получить правильный ответ и оценить правильность сказанного*’. Из толкования видно, что отнесенность к будущему следует из образной составляющей, выделенной в толковании курсивом. Понятно, что внутренняя форма идиомы, хотя и является важной частью плана содержания, менее центральна, чем актуальное значение. Хорошее идеографическое описание должно учитывать подобную неоднородность.

С теоретической точки зрения неоднородность такого рода можно рассматривать как отражение неоднородности содержания более крупной единицы — высказывания, которое разделяется на утверждение, пресуппозиции и следствия, тему и рему, данное и новое и т. п. Такое структурирование необходимо, поскольку содержание сообщения должно быть «упаковано» в линейную структуру. Хотя теоретические основания такого устройства плана содержания языка ясны, в практической лексикографии это не получило должного отражения. В Тезаурусе предпринята попытка разделения таксона на центральную и периферийную части. В центральную часть таксона попадают те идиомы, в плане содержания которых соответствующий семантический компонент является значимым, коммуникативно важным, а в периферийную часть — те идиомы, которые связаны с семантикой данного таксона не прямым образом — через следствия, пресуппозиции, периферийные и факультативные семы.

Применительно к уже разобранному примеру идиомы *поживем — увидим* этот принцип устройства Тезауруса работает таким образом, что она попадает в два таксона Тезауруса: в центральную часть таксона НЕОПРЕДЕЛЕННОСТЬ и в периферию таксона БУДУЩЕЕ. Идея ‘неопределенности’ в семантике рассматриваемой идиомы оказывается важнее, чем идея ‘будущего’.

Разберем еще один подобный пример. Идиома *грести под себя* представлена в центральной части таксона БЕЗРАВСТВЕННОСТЬ, БЕССОВЕСТНОСТЬ; ПОДЛОСТЬ и в периферии таксона ПОЛУЧЕНИЕ МАТЕРИАЛЬНЫХ БЛАГ. Интуитивно очевидно, что при употреблении данной идиомы сема ‘безнрав-

[3] Толкование дается в соответствии с принципами, изложенными в ФОС.

ственность’ стоит в фокусе внимания и тем самым более значима, чем сема ‘получение материальных благ’. Эта интуиция подтверждается введением идиомы в контекст запрета: фраза *Перестань, наконец, грести под себя!* понимается не столько как запрет на получение материальных благ, сколько как требование их адекватного (соответствующего представлениям говорящего) распределения.

Идиома *у Х-а [все] как у людей* попадает в центральную часть таксонов СООТВЕТСТВИЕ СВОЕМУ ОБЩЕСТВЕННОМУ ПОЛОЖЕНИЮ, СТАТУСНОЙ НОРМЕ и СООТВЕТСТВИЕ ЭТАЛОНУ, ОБРАЗЦУ, а также в периферийную часть таксона ОБЕСПЕЧЕННОСТЬ, НЕБЕДНОСТЬ. Это объясняется тем, что конкретная характеристика, по которой Х попадает в множество людей, обозначаемое идиомой, не обязательно связана с материальным ресурсом. Таким образом, сема ‘обеспеченность, небедность’ оказывается одним из возможных семантических следствий.

В некоторых случаях включение идиомы в периферию таксона объясняется факультативностью соответствующей семы. Например, идиома *на вес золота* включена в центральную часть таксона ВАЖНОСТЬ, ЗНАЧИМОСТЬ, ЦЕННОСТЬ и в периферийную часть таксона МАЛО, МЕНЬШАЯ ЧАСТЬ, НЕДОСТАТОЧНО; НЕМНОГО. Целесообразность такого решения следует из значения данной идиомы: ‘кто-л./что-л. рассматривается как большая ценность, часто имеющаяся в недостаточном количестве, *как если бы это был драгоценный металл*’ [ФОС: 60]. В то время как сема ‘большая ценность’ является центральной и не нейтрализуется ни в одном из возможных контекстов, сема ‘недостаточно’, которая вводится компонентом толкования ‘часто’, реализуется далеко не во всех контекстах; ср. (1), где идея недостаточности, возможно, присутствует и (2), где этой идеи нет.

(1) Времена-то, сам знаешь, какие наступают! Литературе вроде как вольную дают… Теперь люди с моим уровнем критического мышления *на вес золота* будут! *[Ю. Поляков. Козленок в молоке]*

(2) В мрачные времена Средневековья карлики ценились *на вес золота*. Всякий мало-мальски уважающий себя правитель должен был иметь при себе хотя бы одного. *[«Московский комсомолец»]*

Встречаются случаи, когда идиомы попадают только в периферийные части таксонов. Это связано с тем, что идиоматика, в отличие от лексики естественного языка, далеко не полностью покрывает все возможные семантические поля. Когда идиома попадает только в периферийную часть того или иного таксона, это означает, что центральный для ее значения семантический компонент оказывается для идиоматики уникальным. Иными словами, для этой идиомы не находится достаточного количества «партнеров», которые могли бы образовать свой отдельный таксон. Например, идиома *язык сломаешь* представлена только в периферии таксонов ГОВОРЕНИЕ и ТРУДНОСТИ, НЕПРИЯТНОСТИ, БЕДА, ПРОБЛЕМА. Это связано с тем, что в центре семантики этой идиомы стоит компонент ‘трудно произнести’, а такого таксона в Тезаурусе не выделилось, поскольку больше нет идиом, обозначающих нечто труднопроизносимое. При этом ближайшими к центральной семе оказываются идеи ‘говорение’ и ‘трудности’.

Использованные в Тезаурусе парадигматические отсылки также призваны отразить неоднородность плана содержания тех или иных идиом (хотя, как уже было сказано, их главная задача — передать сетевую структуру идиоматики). Например, идиомы таксона БОГАТСТВО, МАТЕРИАЛЬНОЕ БЛАГОПОЛУЧИЕ имеют в своей семантической структуре тривиальный компонент ‘много’, поскольку богатство предполагает значительное количество какого-то материального ресурса. В принципе, их следовало бы полностью продублировать в таксоне МНОГО, БОЛЬШАЯ ЧАСТЬ, ДОСТАТОЧНО, СЛИШКОМ МНОГО. И технически и по сути дела это неудобно, поэтому в таксоне МНОГО, БОЛЬШАЯ ЧАСТЬ… содержится отсылка к таксону БОГАТСТВО, МАТЕРИАЛЬНОЕ БЛАГОПОЛУЧИЕ.

***

Попытка практического описания идиоматики с точки зрения структуры семантических полей показывает, что дедуктивный подход оказывается неэффективным. Применение иерархической логической схемы к материалу фразеологии приведет к появлению множества пустых или почти пустых таксонов. С другой стороны, многие достаточно представительные для фразеологии таксоны были бы упущены при дедуктивном подходе.

Индуктивный подход, конечно, не дает единого последовательного иерархического описания, однако использование парадигматических отсылок и разделение таксона на центральную и периферийную части позволяет отразить по крайней мере часть сложных семантических отношений, которые присущи фразеологической системе.

Литература

1. *Апресян Ю. Д.* Новый объяснительный словарь синонимов: концепция и типы информации // Новый объяснительный словарь синонимов русского языка. Под общим руководством Апресяна Ю. Д. М., 1995, 7–118.
2. *Варга Д.* Методика подготовки информационных тезаурусов // Сборник переводов по вопросам информационной теории и практики. № 17. М., 1970.
3. *Падучева Е. В.* Понятие презумпции в лингвистической семантике // Семиотика и информатика. Вып. 8. М.: ВИНИТИ, 1977, с. 91–124.
4. *Тезаурус* — Словарь-тезаурус современной русской идиоматики. М., 2007.
5. *ФОС* — Фразеологический объяснительный словарь русского языка. М., 2009.
6. *Dornseiff F.* Der deutsche Wortschatz nach Sachgruppen. 5. Aufl. Berlin: de Gruyter, 1959.
7. *Hallig R., Wartburg W. von.* Begriffssystem als Grundlage für die Lexikographie. Versuch eines Ordnungsschemas. 2., neu bearb. u. erw. Aufl. Berlin: Akademie-Verlag, 1963.
8. *Quillian M. R.* Semantic memory // Semantic information processing. Cambridge, MA: The MIT Press, 1968.
9. *Roget's* Thesaurus of English words and phrases. New edition. Harlow, 1987.
10. *Wierzbicka A.* Semantics: primes and universals. Oxford/New York: Oxford University Press, 1996.

Еще раз о факторах идиоматичности: тавтология и онимизация

One more about factors of idiomaticity: tautology and onymization

Баранов А. Н. (baranov_anatoly@hotmail.com)
Институт русского языка РАН

В докладе рассматриваются два новых параметра идиоматичности: тавтология и онимизация. Сущность тавтологии заключается в том, что во внутренней формы идиомы повторяются те или иные компоненты актуального значения. Онимизация — это перенос свойств имени собственного на имя нарицательное. В докладе обсуждаются примеры тавтологизации и онимизации как факторов идиоматичности.

1. Факторы идиоматичности

Специфика фразеологии в том, что она формируется неоднословными выражениями, занимающими промежуточное положение между грамматикой и словарем. Действительно, по формальной структуре фразеологизмы должны были бы относиться к ведению грамматики, однако они порождаются по нерегулярным правилам и, соответственно, их экономнее хранить в словаре. Нерегулярность правил формирования и функционирования фразеологизмов проявляется в идиоматичности их семантики.

В ряде исследований было показано, что идиоматичность представляет собой сложный феномен, который определяется тремя важнейшими факторами: переинтерпретацией, непрозрачностью и усложнением способа указания на денотат [Баранов, Добровольский 1996; 2008]. Переинтерпретация в общем случае сводится к переосмыслению значения одной или нескольких форм, входящих в состав идиомы, или к переосмыслению всего выражения в целом. Так, актуальное значение идиомы *перекрыть кислород* — ‘лишить кого-л. возможности действовать как он хочет (часто с целью заставить его делать что-л. другое), *что сопоставляется с прекращением подачи вещества, жизненно необходимого для дыхания*’[1] — возникло в результате переинтерпретации свободного словосочетания *перекрыть кислород*, используемого в контекстах типа *Чтобы прекратить химическую реакцию надо перекрыть кислород*. Непрозрачность связана с наличием во фразеологизме таких выражений, которые отсутствуют в составе словаря современного языка и воспринимаются носителями как что-то непонятное (слово *баклуши* в *бить* ***баклуши*** или форма *зги* в идиоме *не видно не* ***зги***). Усложнение способа указания на денотат как фактор идиоматичности реализуется в таких ситуациях, когда имеется стандартное обозначение некоторой сущности — например, название *Петербург* и альтернативная усложненная номинация *Северная Пальмира* или наречие *навсегда* (как стандартный способ называния) и идиома *на веки вечные* — как усложненный способ выражения того же смысла.

Факторы или аспекты идиоматичности — переинтерпретация, непрозрачность и усложненность способа указания — сами по себе неоднородны и состоят из различных типов [Баранов, Добровольский 1996; 2008]. Так, в сфере фразеологии переинтерпретация в точном смысле — это переосмысление прямого значения соответствующего словосочетания (ср. выражение *верхушка айсберга* в значении «часть реального природного феномена» и в значении «известная часть некоторой ситуации, большая часть которой остается тайной»). Интенсиональная переинтерпретация характеризует такие идиомы, которые в прямом значении не имеют экстенсионала (*буря в стакане воды, ходячий анекдот, показать небо в алмазах*). Среди типов непрозрачности выделяется, например, непрозрачность редуцированной формы (*пуститься во все тяжкие*, вместо *пуститься во все тяжкие грехи*) и компонентная непрозрачность (ср. несловарность компонента

[1] Толкование М. М. Вознесенской из [ФОС]. Здесь и далее используется нотация записи толкования, принятая в [ФОС], в соответствии с которой внутренняя форма идиомы отражается в семантической экспликации курсивом.

лясы в *точить* ***лясы*** или *утряку* в идиоме *по* ***утряку***) Существуют и другие типы указанных факторов идиоматичности, обсуждаемые в [Баранов, Добровольский 1996] и более подробно рассматриваемые в [Баранов, Добровольский 2008]. Однако исчерпанность списка типов переинтерпретации, непрозрачности и усложненности указания вряд ли можно доказать. И действительно, обнаруживаются и другие типы факторов идиоматичности, не упомянутые в указанных исследованиях, два из которых будут рассмотрены ниже. Это тавтология и онимизация.

2. Тавтология как тип усложнения способа указания

Начнем с тавтологии. Она характеризует идиомы *маменькин сынок, маменькина дочка, папина/папенькина дочка, папин/папенькин сынок*. Суть данного фактора идиоматичности в том, что внутренняя форма тавтологичных фразеологизмов повторяет какие-то части плана содержания своих же компонентов — чаще всего главных в смысловом отношении. Действительно, значение слова *сын* (и уменьшительная форма *сынок*) включает информацию о том, что лицо, являющееся сыном, находится в отношении прямого родства с отцом и матерью. Аналогично слово *дочь* (*дочка*) также указывает на родственные отношения с отцом и матерью. Повторение данной семантической информации во внутренней форме явно избыточно — тавтологично, поскольку сын действительно является сыном своей матери и отца, а дочь, конечно же, дочерью своей матери и отца.

Можно показать, что тавтология во внутренней форме не может быть видом переинтерпретации. Во-первых, сами слова *сын* (*сынок*) и *дочь (дочка)* не подвергаются переинтерпретации (переосмыслению). Они и в соответствующих идиомах указывают на определенный тип родства. Во-вторых, в нормальном случае (вне ситуации языковой игры) выражения *мамин сын (сынок), мамина дочь (дочка), папина дочь (дочка), папин сын (сынок)* как свободные словосочетания вообще не используются. То же самое можно сказать об этих выражениях с вариантами *маменькин* и *папенькин*. Казалось бы, здесь можно говорить об интенсиональной переинтерпретации, то есть о случаях типа *буря в стакане воде, адская машина, кровавая баня*. Однако при интенсиональной интерпретации буквальное прочтение невозможно, поскольку соответствующая ситуация в реальности не существует. Между тем, сочетания *мамин сын (сынок), мамина дочь (дочка)* и т. д. в прямом значении обозначали бы реально существующих лиц, однако чрезмерное дублирование смыслов в словах *мама, сын (сынок), дочь (дочка)* приводит к тому, что используется только переносное значение соответствующих фраз. Не подходят и другие случаи переинтерпретации, обсуждаемые в [Баранов, Добровольский 1996; 2008]. Тавтология во внутренней форме не может рассматриваться и как вид непрозрачности, поскольку все слова, входящие в состав идиом *маменькин сынок, маменькина дочка, папина дочь (дочка), папин сын*, являются частью словаря современного русского языка. Иными словами, ничего «непрозрачного», «непонятного» во внутренней форме данных идиом нет.

Остается последний фактор идиоматичности — усложнение способа указания на денотат. И действительно, тавтология, конечно же, усложняет способ указания: вместо *сын (сынок)* мы говорим *мамин/маменькин сынок (сын),* вместо *дочь (дочка) — маменькина/папенькина дочка (дочь*). Похожим — хотя и не таким! — образом усложняется способ указания с помощью квазисинонимов: *всем и каждому, все и вся, иметь место быть, всего ничего* и под.[2] В чем-то похоже на тавтологию и усложнение с помощью редупликаций: *на веки вечные, век вековать, криком кричать*.

3. Сущность идиоматичности при тавтологии

Тавтологизация основывается на повторении каких-то компонентов семантики слов, входящих в состав идиомы. А повторение иконично в том смысле, что, повторяя что-то, мы коммуникативно высвечиваем соответствующий смысл. Именно так и устроена семантика обсуждаемых идиом. Выражение *маменькин сынок* имеет два значения:

1. Мальчик, ведущий себя так, как это свойственно девочкам — вежливый, не участвующий в проказах, драках, получающий хорошие оценки в школе, обучающийся игре на музыкальных инструментах и т. п., *и избыточно с точки зрения информативности названный по родственной принадлежности со стороны матери в скрытом противопоставлении отцу и тем самым осмысляемый как воспитанный матерью.*
2. Мужчина, не проявляющий черт характера, свойственных мужчинам: твердости в принятии решений, настойчивости, способности переносить лишения, страдания и т. п. *и избыточно с точки зрения информативности названный по родственной принадлежности со стороны матери в скрытом противопоставлении отцу и тем самым осмысляемый как воспитанный матерью.*

[2] Отмечу, что усложнение дескрипции с помощью квазисинонимов вполне могло бы быть названо авторами уже упоминавшихся работ «тавтологией», хотя и по другим основаниям.

Первое значение обнаруживается в контекстах типа *Ольга никогда к нему не относилась серьезно: лопоухий соседский мальчишка, мученик пианино, губошлеп* и ***маменькин сынок****; Одна моя знакомая, милая женщина, не желая, чтобы ее сын вырос нежным* ***маменькиным сынком****, разговаривает с ним на уличном сленге*. Второе значение реализуется в примерах, в которых речь идет о взрослом мужчине: *Я, конечно, не Саул, и я тебя понимаю. Никакой ты не враг. Но что* ***мамкин сынок****, это точно.* [В. Дудинцев. Белые одежды]; *На кой хер я ввязался в этот идиотский спор? Ведь мне все-таки не восемнадцать лет, ведь я все-таки не* ***маменькин сынок****.* [В. Аксенов. Московская сага. Тюрьма и мир]. Из толкований хорошо видно, что именно появление компонента *маменькин*, дублирующего часть смысла слова *сынок*, фокусирует внимание на этом смысле и выделяет те свойства характера, которые, проявляясь у мужчины (мальчика — в первом значении), типичны для женщин. Разумеется, в этом коммуникативном высвечивании играет дополнительную роль и форма прилагательного — *маменькин*, вместо *мамин* (впрочем, в текстах присутствует и эта форма).

В отличие от идиомы *маменькин сынок*, ее аналог — идиома *маменькина дочка* — не имеет первого значения, то есть не употребляется применительно к маленьким девочкам. На то есть чисто прагматические основания: девочки в типичном случае воспринимаются как аккуратные, вежливые, добрые, неконфликтные и т. д. Таков, если угодно, фрейм. Идиоматика по большей части фиксирует отклонения от фрейма. Именно поэтому первое значение у идиомы *маменькина дочка* отсутствует. Что касается взрослых женщин, то девочки, видимо, в обычном случае с возрастом портятся — отсюда и необходимость фиксации необычных случаев отклонения от стандарта, которые, впрочем, все равно оцениваются говорящим не слишком положительно:

> *маменькина дочка* = ‘Женщина, утрированно проявляющая черты характера, свойственные именно женщинам: вежливость, аккуратность, терпимость, склонность к музыке, получению образования и т. д., *и избыточно с точки зрения информативности названная по родственной принадлежности со стороны матери в скрытом противопоставлении отцу и тем самым осмысляемая как воспитанная матерью*’.

Идея «утрированного» проявления женских черт характера передает слабую отрицательную оценку, которая часто проявляется в иронии говорящего: *Это была высокая девица с нежными чертами лица, безукоризненной фигурой. Происходила она из известной московской семьи Фидлеров <…>. Ей пришлось тоже немало хлебнуть, этой кисейной барышне,* ***маменькиной дочке****, с ее аристократическими манерами*. [Д. Гранин. Зубр]; *Воспитана, вежлива, но* ***маменькиной дочкой*** *ее не назовешь*. [Корпус Публ.]

В русской идиоматике представлены и два других комбинаторных варианта аналогичного повторения родственных связей во внутренней форме — это идиомы *папина/папенькина дочка* и *папенькин сынок*. Ср. контексты употребления идиомы *папина/папенькина дочка*: *<…> я поняла, что я* ***папина дочка****. Мы все в папу.* [А. Вампилов. Старший сын]; *Она росла настоящей* ***папиной дочкой****, она и внешностью пошла в Вакаров, и характером*. [А. Маринина. Убийца поневоле]. Как и в ранее разобранных случаях, коммуникативное высвечивание родственной связи с отцом приводит к концентрации внимания на сходстве свойств характера отца и дочери, близости их взглядов и т. п.: *папина дочка* = ‘Дочь, которая привязана к отцу больше, чем к матери, похожа на него характером, унаследовала какие-то его способности и т. п. или которой отец уделял больше времени при воспитании[3], *избыточно с точки зрения информативности названная по родственной принадлежности со стороны отца в скрытом противопоставлении матери*’.

Сравнительно редко употребляющаяся идиома *папенькин сынок* высвечивает в актуальном значении несколько неожиданные компоненты семантики по сравнению с аналогичными выражениями *маменькина дочка* или *папина дочка*. В имеющихся контекстах употребления речь идет об использовании сыном протекции или помощи со стороны отца: *Он всунул меня в инструментальный цех — работа непыльная, не надо в мазуте копаться<…>. Только я его предупредил, чтоб на заводе близко ко мне не подходил, не позорил. Я ж не* ***папенькин сынок*** *какой-нибудь*. [В. Козлов. Школа]; — *Распределили, — пожал он плечами. — Обязательно что ли у него под крылом сидеть. Чтобы все кивали — вот, мол,* ***папенькин сынок***. [Р. Гусейнов. Ибо прежнее прошло]; *–Мамочка, не нужно меня искать <…>! Миллионы парней вроде меня едут на фронт. Я не хочу быть маменькиным, а тем более* ***папенькиным сынком****, не хочу позорить отца!* [В. Аксенов. Московская сага. Война и тюрьма]

В последнем примере при перечислении свойств *маменькиного сынка* и *папенькиного сынка* речь идет в первом случае о человеке, не проявляющем черт характера, свойственных мужчинам, а во втором — об использовании связей отца для того, чтобы избежать отправки на фронт в действующую армию. Иными словами, противопоставления между отцом и матерью во внутренней форме в рассматриваемом случае нет. Соответственно, толкование идиомы *папенькин сынок* могло бы выглядеть следующим образом: ‘Относительно молодой мужчина, использующий помощь своего отца (его связи, положение, ста-

[3] Интересно, что интуитивно хочется компонент ‘которой отец уделял больше времени при воспитании’ оставить для данной идиомы в актуальном значении, а в идиоме *маменькин сынок* перенести во внутреннюю форму.

тус и т. п.) для карьерного роста, решения возникающих житейских и др. проблем *и избыточно с точки зрения информативности — названный по родственной принадлежности со стороны отца*'. Интересно, что выражение соответствующего комплекса смыслов по отношению к сыну или дочери, использующих аналогичный «административный» ресурс своей матери, в идиоматике не предусмотрено.

4. Онимизация как фактор идиоматичности

Под онимизацей будем понимать переход имен нарицательных в имена собственные или структурное преобразование имен нарицательных в имена собственные, приводящее к приобретению ими тех или иных семантических свойств имен собственных. Типичные примеры идиом, возникших в результате онимизации — выражения *Софья/Степанида Властьевна* (по отношению к советской власти), *Манька Величкина* 'мания величия'; *Голим Голимыч, Писец Петрович* 'неуспех' и пр.

Онимизация как тип идиоматичности существенно отличается от тавтологии. Во-первых, онимизация чаще всего связана с языковой игрой. Хотя персонификация абстрактного — типичный языковой прием метафоризации, все-таки использование полной формы имени собственного — имени и патронима — по отношению к абстрактной сущности нетривиально. Излишняя уважительность порождает в данном случае смеховой эффект. Во-вторых, онимизация довольно продуктивна и регулярна. Причем регулярна в двух смыслах: из-за регулярности персонификации как способа метафорического осмысления и из-за наличия правила образования отдельной группы выражений, входящих в состав обсуждаемой группы идиом. Сама регулярность образования некоторых выражений такого рода ставит вопрос об отнесении их к идиомам в точном смысле. Выражения *Голим Голимыч* и *Писец Петрович* возникли по общему правилу: «слово, обозначающее проблему или неуспех → слово, обозначающее проблему или неуспех, переосмысленное как имя лица + отчество, образованное от первого слова или напоминающее его[4]». Так, образовав по данному правилу форму *Облом Обломыч*, я обнаружил в Интернете множество примеров употребления этого выражения:

(1) Приезжаем на вокзал, и начинается ***Облом Обломыч***. Та гостиница, в которую мы хотели поселиться, оказывается закрытой, в остальных цены кусаются.

(2) Подруга на меня сердилась две недели — вместо того, чтоб попить холодненького пивка после баньки, она вкушала ***облом обломыча***! С новой приятельницей я напилась.

В отличие от *Облом Обломыча*, выражение *Конец Концович* (опять-таки сконструированное по правилу и лишь позже найденное в Интернете) имеет несколько иную семантику, мотивированную идиомой *в конце концов*:

(3) В самом пресамом ***конце концовиче*** сцена из Вайнеров: герой в детском доме, мальчика хочет усыновить, трогательно, аж тошнит. [о кинофильме Е. Кончаловского «Побег»]

(4) В ***конце-концовиче*** Госдумой был принят Федеральный закон о внесении изменений в статью 4 Федерального закона «О Центральном банке Российской Федерации».

Продуктивность данного способа образования проявляется в его дальнейшем развитии и усложнении — к имени и отчеству может присоединяться фамилия, образованная на основе исходного слова (имени нарицательном), ср. ответную реплику с использованием выражения *Облом Обломыч Обломатов* как реакцию с семантикой удивления и сожаления:

(5) Парень кричит: — Ты согласна??!! Светка: — Да! Да! Я тебя хочу!! Я СОГЛАСНА НА ВСЕ!!!- Светка бросается на кровать, парень к ней и тут из вороха одежды появляется фигура отца со словами:— Стоп, ребята, я не согласен! — Да... ***Облом Обломыч Обломатов***....

Аналогичная форма легко образуется и для выражения *Конец Концович*:

(6) Ведь всему будет венец — / КОНЕЦ КОНЦОВИЧ РАСКОНЕЦ / Так для этого конца / Мы высылаем удальца / Чтоб с концом он подсобил / Побольше пиплов загубил [А. Палантирский. Под впечатлением шизы]

Регулярность и продуктивность данного способа образования заставляет сомневаться в том, что выражения типа *Облом Обломыч*, *Голим Голимыч* и *Писец Петрович* следует относить к идиомам в точном смысле. Этот вопрос требует дальнейшего изучения.

Еще одно специфическое отличие онимизации от тавтологии как типа идиоматичности в том, что если тавтология относится к усложнению способа указания на денотат, то выражения, возникшие в результате онимизации, в рассматриваемом отношении негомогенны. Так, идиома *Слава Кпсс* (переосмысленный лозунг советских времен *Слава КПСС!*) близка к случаям переинтерпретации (переинтерпрета-

[4] Возможен даже вариант, когда используется любое отчество, образованное от простого русского имени, ср. *Писец Иваныч*.

ция грамматических характеристик), идиома *Софья/Степанида Властьевна* соединят черты переинтерпретации и усложнения способа указания на денотат, а *Манька Величкина*, скорее, должна быть отнесена только к усложнению способа указания на денотат.

Такая негомогенность единиц, входящих в класс онимизаций, заставляет предположить, что внутри этой группы находятся различные подклассы, объединяемые только общей идеей преобразования имени нарицательного в имя собственное (вид метафоры персонификации). Действительно, кроме обсужденных примеров, имеются и другие фразеологизмы, возникшие в результате онимизации, количество которых невелико, но семантика весьма разнообразна. Ср. следующие семантические группы:

- имена фольклорных персонажей (*Дед Мороз; Михайло Иваныч; Лиса Патрикеевна*);
- названий растений (*Ванька Мокрый; иван-да-марья*)
- шутливые названия напитков (*Александр Третий* — винтаж тройного одеколона с одеколоном «Саша» — как протест против монополии советского государства на спиртные напитки, особенно при отсутствии денег; *Кровавая Мэри* — вполне современный коктейль, состоящий из водки и томатного сока);
- эвфемистические названия феноменов сферы секса (*играть/жить с Дунькой Кулаковой; половые сношения с Дунькой Кулаковой*);
- обсценные выражения типа *П…. Ивановна/Иванна*; *Х.. Иванович*;
- отсутствие ресурса (*голый Вася*);
- некоторые феномены сферы болезни и смерти (*пришел Кондратий, хватила Кондрашка*).

Теоретический статус онимизации как фактора идиоматичности требует дальнейшего изучения. Это особенно важно с учетом того, что онимизация является активным процессом образования новых форм в современной разговорной речи.

***

Идиоматичность — универсальное свойство языковой системы. Выявленные к настоящему времени факторы идиоматичности (см. [Баранов, Добровольский 1996; 2008]) существенны, но вряд ли исчерпывают все возможные языковые механизмы возникновения форм, идиоматически выражающих смысл. Логически исчерпывающее перечисление здесь вряд ли возможно. Остается путь изучения конкретного материала — в частности, репрезентативных словников идиом. Представленный здесь результат анализа словника «Словаря-тезауруса современной русской идиоматики» [Тезаурус] позволяет с осторожным оптимизмом оценивать перспективность подобных исследований.

Литература

1. *Баранов А. Н., Добровольский Д. О.* Идиоматичность и идиомы // Вопросы языкознания. № 5. 1996.
2. *Баранов А. Н., Добровольский Д. О.* Аспекты теории фразеологии. М., 2008.
3. *Тезаурус* — Словарь-тезаурус современной русской идиоматики. М., 2007.
4. *ФОС* — Фразеологический объяснительный словарь русского языка. М., 2009.

Лингвистический анализ стереотипов: баланс между текстом и смыслом

Linguistic analysis of stereotypes: a balance between texts and meanings

Бергельсон М. Б. (mirabergelson@gmail.com),
Некрасова А. Е. (anna.nekrasova@gmail.com)

МГУ имени М.В. Ломоносова

В статье на материале английских и французских газетных изданий рассматриваются англо-французские этнические стереотипы. Стереотипы описываются как особая разновидность лингвокультурных моделей. Анализируются лингвистические способы подачи и интерпретации информации в дискурсе.

1. Лингвокультурные модели (ЛКМ) как структурированные кванты культурно языковой картины мира являются центральным компонентом процессов интерпретации и понимания: с ними сравнивается новая информация, получаемая участниками коммуникативного взаимодействия в дискурсе. Обратной стороной этого процесса является то, что новая получаемая в дискурсе информация, может дополнять, уточнять, расширять, и видоизменять имеющиеся у коммуникантов ЛКМ. При сравнении новой информации с ЛКМ она может ***соответствовать*** или ***противоречить*** ЛКМ. Другим, независимым от первого, фактором, влияющим на процессы интерпретации, будет *принятие* или *непринятие* этой новой информации участниками коммуникации. Результатом взаимодействия этих факторов являются подкрепление, уточнение, расширение ЛКМ или частичный или полный отказ от ЛКМ (=пересмотр ЛКМ), а также — создание новых ЛКМ.

1.1. В статье делается попытка рассмотреть процессы интерпретации, связанные с такими специфическими ЛКМ, каковыми являются *этнические стереотипы*. Более точно, целью данного этапа исследования является изучение, первоначальная типологизация и классификация дискурсивных стратегий и языковых средств, которыми оперирует автор текста, с целью оказать воздействие на ЛКМ (=этнические стереотипы) адресата.

Поскольку исследователь, занимающийся изучением стереотипов, не имеет доступа к сознанию людей, он вынужден прибегнуть к анализу письменных и устных источников, содержащих стереотипы. Поэтому объектом анализа являются, строго говоря, не сами стереотипы, а высказывания, их отражающие. В связи с этим обстоятельством возникает необходимость лингвистического анализа стереотипов, наряду с более распространенным в отношении стереотипов, анализом культурологическим. Из этого следует, что предметом рассмотрения выступают лингвистические механизмы, реализующие и сопровождающие стереотипные высказывания в дискурсе. Мы опираемся на следующую гипотезу: стереотип является ментальной сущностью, которая, однако, доступна наблюдению и верификации, будучи зафиксированной в дискурсе с помощью определенных языковых средств, приемов и механизмов.

Данная работа посвящена непосредственно поиску, анализу и классификации этих языковых средств. Материалом для исследования послужили статьи из качественной британской и французской прессы. Рассматривались следующие издания: *Daily Telegraph*, *Guardian, Times, Independent, Libération, Figaro*. Методом сплошной выборки были рассмотрены статьи из названных шести печатных изданий за два года; общая выборка составила более 192 тысяч слов. Были проанализированы 469 случаев встречаемости англо-французских этнических стереотипов в 235 текстах.

1.2. Ввиду разнообразия определения понятия стереотип в качестве рабочего принимается

следующее определение: (этнический) стереотип — это лингвокультурная модель, распространенная в некоторой лингвокультурной среде, фиксирующая представления данной среды о той или иной группе (нации) — в частности, заостряя и упрощая некоторые ее черты — и позволяющая апеллировать к этим представлениям в дискурсе. Удобство этого определения состоит в том, что оно позволяет взглянуть на стереотип не просто как на некое обобщенное представление одной нации о другой, но как на множество структурированных фрагментов когнитивной модели. А процессы внедрения стереотипов или закрепления их в сознании могут описаны как взаимодействие информации, извлекаемой из текста, с имеющимися в голове у читателя ЛКМ с целью подкрепления или дополнения (расширения, уточнения) последних.

2. Имплицитные и эксплицитные способы воздействия

Лингвистические приемы, с которыми связано выражение стереотипов в дискурсе, делятся на две большие категории. Это приемы семантические и приемы стилистические. Семантические приемы непосредственно отражают и создают стереотипы в дискурсе. Стилистические приемы украшают стереотипы, выраженные другими языковыми средствами, делают описываемый стереотип ярче и рельефнее.

Не вся информация, вкладываемая автором в дискурс, бывает выражена эксплицитно. Часть имплицитно выраженной информации передается автором с помощью дискурсивных тактик, таких как подача информации в пресуппозиции и с помощью инференций. Поскольку косвенность кодирования, характерная для имплицитной информации, обычно вызывает ассоциации с меньшей значимостью, имплицитная информация играет большую роль в манипулировании сознанием и передаче информации, которую трудно или невозможно логически обосновать, что особенно актуально при передаче стереотипов.

2.1. Можно предложить следующую классификацию высказываний, отражающих стереотипы, по степени эксплицитности содержащейся в них культурологической информации:

- стереотипы выражены эксплицитно;
- стереотипы поданы в пресуппозиции;
- стереотипы подаются как инференции;
- стереотипы передаются при помощи приема переключения кодов (code-switching)
- стереотипы передаются через описание ситуации действительности, внешней по отношению к данному коммуникативному взаимодействию (этот прием может реализовывать разную степень эксплицитности: он может в качестве примера сопровождать эксплицитное называние стереотипа, а может представлять собой и косвенную подачу стереотипа, без его дополнительного эксплицирования).

Остановимся подробнее на приемах пресуппозиции и инференции, а также приеме переключения кодов..

2.1.1. Для определения пресуппозиции воспользуемся формулировкой М.А. Кронгауза: «пресуппозицией называется такой пропозициональный элемент высказывания, ложность которого делает все высказывание неуместным (аномальным)» (Кронгауз 2005: 204). Рассмотрим следующий пример:

*Penniless, he nevertheless had an idea — informed by his **native French** chic — about wrapping the body in a length of fabric …. (26)*

Приведенный отрывок является яркой иллюстрацией введения стереотипа в пресуппозиции. Если данное предложение поставить в отрицательную форму (*he wasn't informed by his native French chic*), сохранится пресуппозиция о том, что французы обладают особым шиком (*there is native French chic*).

Термин «инференция» используется многими исследователями проблем дискурса. Согласно М. Л. Макарову, «инференции — это широкий класс когнитивных операций, в ходе которых и слушающим, и интерпретаторам дискурса, лишенным непосредственного доступа к процессам порождения речи в голове или «душе» говорящего, приходится «додумывать за него» (Макаров 2003: 124). Трудно исчерпывающе ответить на вопрос о том, какой именно лингвистический материал подталкивает адресата к инференциям, так как возможность использовать инференции часто опирается на текущие характеристики дискурсивного события. Однако можно выделить классы языковых единиц (или обозначить приемы), которые как бы «наводят» адресата на необходимость логических операций для выявления скрытых смыслов. Среди них таким свойством обладают слова и словосочетания, являющиеся характеристиками событий (*no longer, unusually, typiquement*), а также частицы и коннекторы (*therefore, même* и др.).

Иллюстрацией в нашем материале может служить название статьи английского автора Джона Хенли (Jon Henley) "*French no longer bon vivants*". В данном случае словосочетание *no longer* приводит к инференции, с помощью которой восстанавливается пропозиция "French used to be bon vivants".

2.1.2. Прием переключения кодов в языке СМИ не всегда используется с той же целью, что и в художественной литературе, где он часто служит созданию речевого портрета героя. Если говорить об употреблении данного при-

ема в медийном дискурсе, на первый план выходит создание не столько выразительного, или даже комического эффекта, сколько привлечение внимания читателя элементом неожиданности такой языковой формы. В этом случае употребление иноязычного слова в дискурсе в том числе дает понять читателю, что упомянутое явление характерно для той или иной лингвокультуры. Так употребление французского слова в английском тексте позволяет, не прибегая к дополнительным лексическим средствам, ненавязчиво сообщить что упомянутое явление характерно для французской лингвокультуры. Эта ненавязчивость особенно важна для имплицитного введения и поддержания этнических стереотипов. Часто это используется в заголовках статей, например: *"How to tell a guy from un garçon: one's not afraid to dress like his father"* (в связи со стереотипом о шике, с которым одеваются французы).

Покажем использование данной языковой стратегии на примере отрывка из статьи «*Info-intox: l'anglais comme on le "speak" ou on le "love"*», опубликованной в Figaro: *En ce qui concerne les performances amoureuses, Arthur se montre très British. A la classique accusation selon laquelle les Britanniques seraient des amants pitoyables, il répond: «Je ne le pense pas, malgré ma modeste expérience.» Vous avez dit self depreciation (autodérision) ?*

Иноязычное слово в медийном дискурсе не только придает повествованию колорит и необычность, но также позволяет указать на стереотип, не прибегая к излишним словам. Недоговоренная фраза *Вы говорите самобичевание*?, в которой слово *самобичевание* дается на английском языке, позволяет автору статьи, не прибегая к дополнительным лексическим средствам, передать информацию о том, что описываемая черта является характерной для обыденной лингвокультурной картины мира англичан.

3. Языковые приемы и стратегии создания и поддержания стереотипа

Если представленные в предыдущем разделе приемы пресуппозитивной и инференционной подачи смысла позволяют создать и ввести в текст стереотип, то здесь мы рассмотрим некоторые механизмы, которые могут быть условно названы стилистическими приемами. Они служат не выражению стереотипа в тексте, а его более выразительному и тем самым более убедительному звучанию. Если семантические приемы позволяют реализовать стереотип, то стилистические приемы помогают сделать его запоминающимся.

В рамках настоящей работы и семантические, и стилистические приемы рассматриваются в перспективе их функций как дискурсивных стратегий и могут анализироваться с точки зрения трех разных составляющих языковой структуры: с точки зрения использования конкретных лексических средств, с точки зрения использования определенных грамматических конструкций, с точки зрения собственно дискурсивных приемов — как на уровне микро-, так и на уровне макроструктутры дискурса.

С точки зрения лексики речь может идти о метафоре, употреблении оценочной лексики, употреблении иноязычного слова, оксюморона, обыгрывании идиом. С точки зрения использования той или иной грамматической конструкции можно говорить об использовании противопоставлений, повторов, обыгрывании прецедентных фраз.

На уровне дискурса (т.е. текста всей статьи или ее части) выделяются следующие стратегии оформления стереотипов, связанные с семантикой развернутых элементов или организацией всего текста:

- определенные вкрапления в структуру нарратива («лишние» элементы с точки зрения связности повествования и экономии средств);
- ирония;
- прием парадокса/эффекта обманутых ожиданий (например, когда конец статьи противоречит началу);
- ситуация, когда выражение мнения о другой лингвокультурной группе невольно раскрывает ситуацию в «своей» лингвокультуре;
- косвенная подача стереотипа: подбор ситуации, которая являлась бы иллюстрацией того или иного стереотипа (подача стереотипа без слов)

3.1. «Лишние элементы» в структуре нарратива

Говоря о стратегиях оформления стереотипа на уровне целого дискурса, в качестве одного из средств активирования стереотипов в газетном тексте остановимся подробнее на приеме выделения в структуре нарратива элементов, которые являются лишними с точки зрения связности повествования и экономии средств. Кажущаяся «ненужность» подобных элементов делает их ключевыми при дискурсивном анализе. Несущественные с формальной точки зрения (с точки зрения связности), они оказываются важными с идеологической точки зрения (то есть с точки зрения внедрения и распространения стереотипов).

Данную стратегию лингвистического оформления стереотипов можно проиллюстрировать на примере отрывка из статьи E. Brockes «Twenty miles away, and a world apart», опубликованной в *Guardian 5.04. 2004*.

Two women stand before me in the bus queue. The driver asks them in French for the fare of €1.50, ***just over***

***£1**. "Wa'ss that in English?" snaps one. **With the tiniest pause**, the driver replies, "£2," and **politely relieves them of their money**.*

В данном случае стереотип о ловкости и изворотливости французов, существующий в культурно-языковой модели мира англичан, подается имплицитно, через описание некоторой ситуации. Этому способствуют следующие слова и словосочетания: *just over £1, With the tiniest pause* и *relieves them of their money*. Данный мини-рассказ состоит из трех частей — дескриптивной (первое предложение), собственно нарративной и элементов: *just over £1* и *With the tiniest pause,* которые не являются необходимыми относительно обычной структуры рассказа. С точки зрения связности повествования они совершенно излишни: шофер попросил оплатить проезд, женщины спросили, сколько это будет по-английски, он ответил и взял деньги. Но тем не менее эти элементы присутствуют в рассказе. В дискурсе не бывает случайностей или «лишних частей», язык построен по принципу экономии усилий, и в рассказе каждый элемент языковой формы несет определенную смысловую и/или коммуникативную нагрузку. Поэтому фразы *just over £1, With the tiniest pause,* которые с точки зрения структуры нарратива являются лишними, на самом деле выполняют определенную функцию. Именно необязательностью своего присутствия они побуждают читателя задуматься, а зачем эта информация автором здесь сообщается, почему именно в этом месте рассказа (уточнение о том, чему именно соответствует 1,5 евро), какие из этого нужно сделать выводы или оценки. Той же цели служит и фраза *relieves them of their money* — букв. 'освобождает их от денег'. В распоряжении автора существует много способов выразить одну и ту же мысль. То, что он выбирает именно данный способ выражения, а не более нейтральное словосочетание «взять деньги», тоже служит созданию у читателя более яркого образа, определенного мнения, подтверждает существующий стереотип.

3.2. Косвенная подача стереотипа: подбор ситуации, которая являлась бы иллюстрацией того или иного стереотипа (подача стереотипа без слов)

Этот прием подачи стереотипа также просматривается на уровне лингвистической организации выше, чем уровень предложения. Автор вместо того, чтобы прямо назвать стереотип, подает стереотип косвенно. Показывается ситуация, из которой читателю предлагается самому сделать вывод. Такой способ подачи стереотипов идеологически очень удобен. С одной стороны, получается, что автор будто бы придерживается нейтральной позиции, не стремится навязать читателю то или иное мнение, предлагает читателю самому сделать вывод. С другой же стороны, исходя из принципа прагматической мотивированности (Бергельсон 2007), одну и ту же ситуацию можно подать под разным углом зрения, в зависимости от коммуникативных намерений автора. Поэтому выбранный автором статьи способ описания ситуации, скорее всего, повлечет за собой ту интерпретацию, которая удобна автору дискурса. Таким образом, происходит имплицитное насаждение стереотипа, реализуется идеологическая функция медиа дискурса.

В качестве иллюстрации можно привести пример из статьи N. Barley "Sarkozy, c'est Mr Bean", опубликованной в газете *Libération* 28.04.2007:

On fait la queue pour le pain dans les quartiers riches de Londres. Famine ? Pas du tout. C'est qu'une boulangerie française vient d'ouvrir.

Сама по себе описываемая ситуация кажется абсурдной: времена больших очередей за продуктами в Западной Европе давно миновали, особенно в богатых районах английской столицы. Автор пользуется приемом риторического вопроса, совмещенного с гиперболой (предположение, что большие очереди были вызваны голодом). При этом сразу же следует ответ на данный вопрос: причина в открытии новой французской булочной. Читателю предлагается самому домыслить, почему очереди у французской булочной в Лондоне сравнимы с теми очередями, которые были голодные годы? Видимо, ответ кроется в особенных качествах французского хлеба, которых не хватает английскому хлебу. Данный пример является частной реализацией стереотипа о превосходстве французской кухни и неумении англичан хорошо готовить.

3.3. Прием парадокса/эффекта обманутых ожиданий

К приемам, основанным на той же стратегии сопоставления с «нормальным» или «правильным» можно отнести и «обманутые ожидания», когда информация выдаваемая в конце статьи опровергает начало. Так, в начале статьи о телевидении говорится: *From a French perspective, British television is a kingdom of excellence*. Затем в конце: следует фраза: *At the end of the programme, you might win the house if you can answer the following question correctly: which one of these is a traditional Spanish drink? A) sangria or B) shandy.* Это заставляет читателя произвести инференцию. Любому человеку, обладающему хотя бы элементарными знаниями об английской или испанской культуре (англичане, без сомнения, знакомы со своей культурой на таком уровне), ясно, что шенди — напиток английский, а сангрия — испанский. Шенди — это напиток, получаемый при смешении пива и лимонада или имбирного пива (см. *Cambridge Dictionary on-line*). Если провести аналогию с российскими реалиями, то данный вопрос будет звучать примерно

так: «Какой напиток является традиционно русским: ерш или сангрия?» Поэтому величина приза (возможность выиграть дом) в сравнении со сложностью вопроса (элементарно) вызывает юмористический эффект и заставляет читателя задуматься о том, что интеллектуальный уровень викторин, проводимых на британском телевидении, является (если не всегда, то по крайней мере часто) очень низким.

Данная инференция, приводящая к обманутым ожиданиям относительно уровня якобы интеллектуального британского телевидения, в свою очередь, становится ступенькой для создания или подкрепления стереотипов о британцах по принципу «сам дурак».

4. Заключение

Стереотипы представляют собой удобный, хотя и опасный, инструмент освоения действительности. Наряду с другими категориями культурно-языковой картины мира они составляют тот фон, тот «задник» коммуникативного процесса, который необходим для усвоения новой информации. Но помимо информационного обмена важнейшей частью коммуникативных взаимодействий является собственно *воздействие*, реализуемое как подкрепление распространенных в данном дискурсивном сообществе взглядов и установок. Подкрепление подразумевает, что информация не должна быть отвергнута ни как слишком новая и противоречащая уже известному, ни как слишком хорошо известная, то есть лишенная интереса новизны для адресата. Именно этим цели достигаются использованием пресуппозиций и инференций, соответственно. Другие, так называемые стилистические, приемы апеллирования к стереотипам используют тот же принцип: вводимая информация должна обратить на себя внимание читателя и заставить его сделать свой собственный вывод.

Литература

1. *Бергельсон М. Б.* Прагматическая и социокультурная мотивированность языковой формы // М.: Университетская книга, 2007.
2. *Кронгауз М. А.* Семантика // М.: Издательский центр «Академия», 2005.
3. *Макаров М. Л.* Основы теории дискурса // М.: ИТДГК «Гнозис», 2003.

О корпусе текстов живой речи: новые поступления и первые результаты исследования[1]

The corpus of spoken russian: new receipts and the first results of research

Богданова Н. В. (nvbogdanova_2005@mail.ru)

Филологический факультет Санкт-Петербургского государственного университета; Санкт-Петербург, Россия

В докладе представлены новые поступления в сбалансированную часть Звукового корпуса русского языка — массива текстов живой монологической речи, объединенных едиными лингвистическими, социолингвистическими и психолингвистическими параметрами. Описываются новые блоки такого корпуса и первые результаты исследования его материала.

Корпус текстов живой монологической речи (текстотека бытовых монологов) представляет собой один из блоков Звукового корпуса русского языка (ЗКРЯ), работа над которым ведется на филологическом факультете СПбГУ. Отличительной особенностью этого блока является достаточно строго сбалансированный характер его формирования. Принцип, положенный в основу организации корпуса, условно можно назвать «*принципом ковчега*» («каждой твари по паре»): балансировке подвергается и состав информантов (с точки зрения их социальных и психологических характеристик), и та лингвистическая программа, по которой осуществляется запись их речи (чтение и пересказ двух текстов, описание двух изображений и свободный рассказ — 7 текстов от каждого информанта) (подробнее см.: *Богданова и др.* 2008).

Работа над этой частью корпуса продолжается в настоящее время по двум направлениям: пополнение его новыми текстами и анализ существующего материала. Целью настоящего сообщения является как раз представление новых структурных частей корпуса и первые конкретные результаты его анализа.

На сегодняшний день структура корпуса выглядит следующим образом — см. таблицу 1:

По сравнению с предыдущим составом, текстотека пополнилась следующим образом:

- два блока интерферированной русской речи иностранцев — американцев и китайцев. Подобный материал дает возможность сравнивать специфические черты спонтанной русской речи говорящих на родном и неродном языке, выявлять универсальные черты спонтанности (в рамках различных коммуникативных сценариев)

Таблица 1. Сбалансированный материал Звукового корпуса русского языка

MED	Речь медицинских работников	32 диктора, 210 текстов	6 часов звучания
JUR	Речь юристов	40 дикторов, 322 текста	16 часов звучания
RKI	Речь преподавателей РКИ	20 дикторов, 70 текстов	3,5 часа звучания
STUD	Речь студентов	5 разных блоков	7 часов звучания
COMP	Речь «компьютерщиков»	12 дикторов, 32 текста	72 мин звучания
RIA	Интерферированная русская речь американцев	68 дикторов, 204 текста	10 часов звучания
RIK	Интерферированная русская речь китайцев	2 диктора, 4 текста	10 минут звучания

[1] Исследование выполнено при поддержке гранта РФФИ «Изучение зависимости речевых характеристик от условий коммуникации (корпусное исследование на материале повседневной русской речи)» (проект 10-06-00300).

и те, что обусловлены интерференционными процессами, возникающими при контакте двух конкретных языков. Насколько можно судить, межъязыковая интерференция на уровне спонтанного речепроизводства вообще редко становилась предметом лингвистического анализа;

- блок профессионально окрашенной бытовой речи «компьютерщиков» (программисты, системные администраторы, преподаватели информатики, студенты, обучающиеся по соответствующим специальностям). Привлечение подобного материала расширяет возможности описания одного из типов внутриязыковой интерференции — между литературной и профессиональной речью носителей языка. В рамках настоящего корпуса это дает возможность сравнивать специфику построения бытового спонтанного монолога в рамках того или иного коммуникативного сценария информантами из разных профессиональных групп: юристами, медиками, преподавателями русского языка как иностранного (РКИ) и «компьютерщиками»;
- небольшой самостоятельный блок речи студентов — точнее, одного информанта, записавшего в течение трех лет полную лингвистическую программу четырехкратно, с месячным интервалом между записями одного и того же коммуникативного сценария (всего 28 текстов). Такой подход позволяет выявить устойчивые черты речевого портрета говорящего, не зависящие от условий протекания конкретного речевого акта, а также степень влияния этих условий на речевую продукцию человека.

Еще одна возможность расширения данной части звукового корпуса и в целом развития исследований в этом направлении видится в пересечении материала двух блоков: данного сбалансированного и корпуса повседневной речи носителей языка «Один речевой день» (ОРД), построенного, в отличие от первого, по «принципу невода» (см. о нем подробнее: *Асиновский и др*. 2008, 2009, а также статью того же авторского коллектива в настоящем сборнике).

Так, возможность сравнить речь человека в разных условиях записи предоставляет эксперимент, осуществляемый в настоящее время: информант-студент, речь которого записывалась многократно в течение трех лет, по одной и той же лингвистической программе (см. выше), записал еще и свой речевой день. Даже с учетом того факта, что он сам прожил этот день «с диктофоном на шее», т. е. знал о проводящейся записи, условия этой записи намного более естественны по сравнению с речью в микрофон. Наши наблюдения показывают, что информант, записывающий свой речевой день, помнит о диктофоне не более первых двух часов, а дальше ведет себя совершенно естественно. В то время как запись с микрофоном и присутствующим экспериментатором ни на секунду не дает информанту расслабиться и забыть о том, что его речь не только слушают, но и фиксируют для дальнейшего анализа. В перспективах этого направления исследования — повторные (как минимум, еще одна, через год) записи речевого дня данного информанта, для выявления всех особенностей его речевого поведения и создания его речевого портрета.

Что касается конкретных результатов исследования материалов этой части ЗКРЯ, то они весьма любопытны и связаны с самыми разными уровнями анализа.

Так, было проведено исследование спонтанной речи информантов из двух профессиональных групп — медиков и юристов, с целью выявления влияния профессии на их бытовую речь. Оказалось, что в обоих случаях это влияние имеет место, но оно принципиально различно. Медиков «выдает» пристрастие к медицинской терминологии, они прибегают к ней тем чаще, чем выше уровень речевой компетенции (УРК) информанта. Более того, информанты с высоким УРК (врачи-преподаватели, профессионально связанные с речью) используют (в рассказе об отдыхе!) достаточно редкие и узко специальные термины (*атеросклероз сосудов головного мозга, сердечная недостаточность, энергозатраты, генотип*), в то время как информанты с низким УРК (медсестры) ограничиваются в своих рассказах терминами, весьма частотными в речи носителей русского языка (*боль, больной, медсестра*). Крайним проявлением влияния профессии на бытовую речь человека можно считать полное переключение говорящего (в рассказе об отдыхе!) на медицинскую тему. Таких случаев в материале выявлено 5 на 30 текстов, при этом ни одного — в монологах информантов с низким УРК. Порой информанты сами осознают эту смену тематики и даже предупреждают об этом собеседника (экспериментатора):

*если бы не Макдоналдс я бы там наверное не выжила потому что японцы едят совершенно изумительную пищу // **э-э опять-таки медицина** // **ну ха-ха вам придется на это сделать сноску** // э-э они где-то лет двадцать назад стали вымирать от инсультов потому что у них произошла американизация и появилась американская пища сеть Макдоналдс все вот эти Кэрролс и так далее // японский организм оказался к этому непривычным и они стали / э-э стали болеть атеросклерозом причем вот именно японская нация стала болеть атеросклерозом сосудов головного мозга // поэтому количество инсультов превзошло превзошло все возможные ожидания они увеличились на порядок по отношению к первоначальному / после этого они вернулись к традиционной пище // традиционную пищу есть могут только японцы* [Инф. 20, жен., высокий УРК].

В целом полученные цифры невелики — около 1 % медицинских терминов на весь лексиче-

ский объем свободных рассказов 30 информантов (4 670 фонетических слов). Однако тема рассказа — о способе проведения свободного времени — вовсе не предполагала использования медицинской терминологии. И взятые для сравнения свободные рассказы информантов-юристов на ту же тему выявили на порядок меньше медицинских терминов (0,07 % от общего числа фонетических слов в текстах), и только широко распространенных — *организм*, *самочувствие*, *реанимация*, *хондроз*, *здоровье*, *инстинкт*, *ангина*. Ни одного случая полного переключения на медицинскую тему в монологах юристов, разумеется, не выявлено.

Влияние профессиональной речи юристов на их бытовую речь оказалось совсем иного рода. Здесь скорее можно говорить о «стилевой разноголосице», о проникновении элементов официально-делового стиля, столь характерного для юридического языка, в бытовые рассказы об отдыхе, ср.:

- *отдых в выходные связан / непосредственно с нахождением дома*;
- *основополагающим моментом моих выходных это должно быть конечно хорошая погода*;
- *их было достаточно много лиц / которые говорили / что нам холодно*;
- *может где-то даже ∫ процентов семьдесят / по моим ∫ э э подсчетам / и наблюдениям / таким образом / отдыхают от ∫ рабочих будней насколько мне известно / больше половины / процентов от всего населения <…> таким образом отдыхают от рабочих будней*;
- *сам процесс отдыха основная составляющая которого / заключается в том что кататься на лыжах*;
- *отдыхаем как совместно с ребёнком / так и не совместно с ребёнком.*

Смешение стилей зачастую рождало в монологах юристов даже комический эффект:

- *лёгкий просмотр телевизора с приёмом завтрака;*
- *долгое / времяпрепровождение вечером с друзьями / в баре;*
- *неприятные ощущения в виде ожогов;*
- *горнолыжный курорт / с очень приятной компанией / с приятным местонахождением;*
- *применяет на себе там / эти косметические средства;*
- *иногда даже / ну употребляется некое количество алкоголя;*
- *в индивидуальном порядке в номере занимаюсь иногда / лечебной такой гимнастикой*.

Два перцептивных эксперимента (на орфографических расшифровках материала и на звучащей речи) показали, что носители русского языка (эксперты-филологи, студенты и преподаватели, 20 чел.) довольно хорошо чувствуют специфику бытовой речи юристов и с высокой степенью вероятности (78–89 % при чтении расшифрованных отрывков и 73–69 % при аудировании) относят предложенные им фрагменты монологов именно к речи юристов (подробнее об этом см.: *Иванова* 2008, 2010). Правда, столь же высок процент отнесения этих фрагментов к профессии менеджера, что позволяет говорить о заметной специфике бытовой речи носителей языка, связанных с деловой сферой жизни.

В ходе анализа материала всего корпуса спонтанных монологов удалось подтвердить справедливость введения в научный обиход такого социологического параметра, как *профессиональное/непрофессиональное отношение говорящего к языку/речи*, и связанного с ним признака *уровня речевой компетенции* говорящего. Анализ речи медиков показал, что даже словарный запас информантов с разным УРК весьма существенно различается: на высоком уровне он составляет около 4 тыс. лексем, на среднем — 3 тыс., на низком — 2 тыс. Другими маркерами УРК говорящего (и все они получили достаточно весомое подтверждение в ходе анализа материала разных блоков корпуса) оказались следующие:

- *степень членимости* текста на единицы, соотносимые с предложением: относительная легкость членения маркирует высокий УРК; наиболее эффективно данный признак диагностирует УРК говорящего на трудных коммуникативных сценариях — в пересказах и описаниях (см. подробнее: *Бродт* 2007);
- *средняя длина «предложений» в словах*: высокому УРК соответствуют длинные «предложения», низкому — короткие (см. там же);
- *употребление вставных конструкций* классического типа (см. *Богданова* 2010б);
- *разнообразие синтаксических конструкций*: диагностирующим УРК является количество сложных «предложений», доля безличных и инфинитивных конструкций и ряд других;
- *разнообразие заполнения синтаксических позиций*, в частности функции инфинитива и употребление причастий и деепричастий (подробнее о маркерах УРК говорящего см.: *Богданова* 2010а).

На обширном материале удалось показать, что синтаксическое и интонационное членение спонтанных текстов разных типов соответствуют друг другу только на 54 % (см. подробнее: *Степихов* 2005).

Анализ *чтения* (корпус речи студентов), несмотря на высокую степень его лингвистической мотивированности первичным текстом, позволил с уверенностью отнести этот вид монолога к разновидностям спонтанной речи. При неподготовленном чтении в речи информантов возникают паузы хезитации, самоперебивы, повторы и различные ошибки — т. е. черты, свойственные любой устной речи:

- *Берестов выехал прогуляться верхом / на всякой случай взяв с собою пары* **<…>** *т***<…>***ри борзых;*
- *огражденный своим чином токмо <…> токмо от побоев; чьи это <…> чьи это дрожки?*
- *и рассказал все / что случи<…> рассказал все / что случилось*; *но он наехал на Берестова вовсе неожи<…> вовсе неожиданно.*

Более того, оказалось, что любое чтение, как неподготовленное, так и подготовленное (разумеется, в рамках лингвистического эксперимента), обладает полным набором черт спонтанности, которые зависят от индивидуальных характеристик говорящего в той же мере, как и любая другая речевая продукция (см. об этом подробнее: *Сапунова* 2009).

На материале монологов-*описаний* (корпус речи студентов) выяснилось, в частности, что в понимании говорящего задача описания изображения существенно расширяется за счет того, что так или иначе связано с изображением: его автор, жанр, искусство в целом, даже собственный опыт человека (подробнее о монологах-описаниях см.: *Филиппова* 2010). В результате отчетливо выделились три класса сценариев:

1) собственно описание:

А) называние, перечисление объектов изображения или событий виртуального мира:
- *летний пейзаж идет тропинка цветочки* [несюж., девушка-нефилолог];

Б) установление отношений объектов или событий с внетекстовой реальностью; суждения, домыслы, догадки говорящего:
- *на шестой* [картинке] *какой-то дядька пожарник судя по всему снимает ее с лестницей* [сюж., юноша-филолог];

2) метакоммуникация — речь информанта о собственной речи («текст о тексте») или о самой ситуации общения:
- *я много не умею говорить / могу красиво молчать* [несюж., юноша-нефилолог];
- *что бы еще вам такого описать* [несюж., юноша-нефилолог].

3) комментирование — речь информанта о собственном опыте, об ассоциациях и т. п.:
- *но он / у нас пейзажист был* [о Шишкине] *так что в общем что с него взять* [несюж., девушка-нефилолог];
- *вообще мне картины Шишкина очень нравятся / еще помню спор был сколько там на картине / сколько там мишек было* [несюж., юноша-филолог].

Пересказы информантов-юристов также позволили выявить различные коммуникативные стратегии, используемые в репродуцированных текстах говорящими с разными социальными и психологическими характеристиками (подробнее о монологах-пересказах см.: *Куканова* 2009).

Я-наблюдатель (наиболее распространенная стратегия — 63 % порожденных текстов, более свойственна для говорящих с высоким УРК, интровертов) — стремление информанта изложить более или менее точно своими словами содержание прочитанного отрывка, передав в сюжетном тексте событийную линию, а в несюжетном — свойства и состояния объектов, которые описываются в первичном тексте:
- *был тёплый июльский день // солнце поднималось ∫ на восходе // солнце было не ∫ не яркое / не огненно-рыжее / оно было светлое и лучезарное // лучи его поднимаясь над землей / пронизывали* (…) *светлые / белые облака // облака <вздох> стоящие над землей* (…) *вверху на небе / лежали белыми белыми яркими ∫ снопами* [Текст 2. Инф. 12].

Я-читатель (32 % проанализированных текстов, самая простая стратегия, в наибольшей степени представлена в монологах говорящих с низким УРК) — стремление информанта передать тематическое содержание первичного текста с выделением главной (доминирующей) темы и ее компонентов:
- *ну / начинается с того что* […] *с описания ∫ (э-э) женщины и пса // входят они / оказываются в какой-то комнате / и реакция пса / вот на ту ситуацию которую он видит // там / какая-то лестница темная* [Текст 1. Инф. 16].

Я-исследователь — (наименее востребованный способ организации текста-пересказа — около 5 %, характерен для мужчин младшей возрастной группы с высоким УРК, экстравертов) попытка информанта проанализировать первичный текст:
- *в общем-то ∫ в данном тексте говорится о ∫ (а-а) ∫ как мы понимаем / о женщине / которая привела собаку / в некую комнату // при этом ∫ (а-а) параллельно видя ∫ (э-э) ситуацию глазами ∫ (а-а) стороннего наблюдателя и глазами собаки* [Текст 1. Инф. 14].

На материале репродуцированных текстов удалось выявить также способы лексического наполнения вторичного текста в сравнении с первичным. *Эндоединицы* полностью повторяют единицы предтекста, а *эзоединицы* — это различные новые единицы, появившиеся только в пересказе, но мотивированные исходным текстом тем или иным способом. Среди последних выявлены, например грамматические и семантические трансформеры (*пощелкала пальцами — пощелкав пальцами, сладкий — сладковатый*), транспозиты (*кричал — крики, солнце — солнечный*), синонимы (*синева — лазурь, земледелец — хлебопашец*), антонимы (*сырость — сухость*), гипонимы/гиперонимы (*пес — животное, комната — помещение*), конверсивы (*она <…> наполнила комнату запахом — комната наполнилась*

запахом), свернутые/развернутые номинации (*личность мужского пола — мужчина, облака с белыми краями — края [облаков] были белые*), ассоциативные замены разного типа (*касторка — валерьянка, небосклону — небоскребу, осколки — обломки, стекло + дверь — окно*), вводящие элементы текста (*повествуется*, *описывается, я прочитал отрывок из Тургенева*), слова, выражающие эмоции (*бедолага пес*) или модальную оценку (*скажем так пожалуй, во-первых, во-вторых*), выполняющие метакоммуникативную (*да сложновато так, насколько я помню, да неправильно*) или дискурсивную функцию (*всё* — в конце монолога), и т. п.

Корпусный характер организации материала позволяет с помощью специальных программ создавать *конкордансы* (алфавитно-частотные словники использованных информантами лексических единиц) разного типа: как общие по разным типам текстов, так и по отдельным группам информантов. По этим словникам можно видеть высокую употребительность таких специфических для спонтанной речи не вполне речевых элементов звуковой цепи, как хезитативы *э-э* или *а-а*, или частиц *вот* и *ну*, также зачастую выполняющих в спонтанной речи не служебную, а чисто хезитационную функцию. Знаменательные части речи уступают таким элементам по частоте употребления порой в десятки раз. Кроме того, появляется возможность сравнивать частотность в устной речи различных грамматических форм одного и того же слова и видеть, например, преобладание начальной формы существительного над косвенными и, наоборот, существенное преобладание личных форм глагола над исходной (инфинитивом). Можно сравнивать между собой и лексикон разных групп носителей языка (организованных по гендерному, возрастному, профессиональному, психологическому и т. п. признакам) в сходных коммуникативных условиях и решать еще множество других исследовательских задач.

Так, в монологах-описаниях среди самых частых слов в словниках практически всех групп информантов (студенты — филологи и нефилологи, юноши и девушки) — отрицательная частица *не*, в какой-то степени свидетельствующая о своеобразной «борьбе» говорящего с трудным коммуникативным сценарием. Значительная часть контекстов с этим *не* — конструкция *не знаю*. А в описаниях несюжетного изображения (самый трудный сценарий) в «верхушку» частотников попало и слово *наверное*, также отражающее поиск говорящими нужного слова, наряду с крайне частотными хезитационными словечками *ну*, *вот*, *это*, *там* и под. Единственным полноценным словом в этой части словника оказалось наречие *очень*, использованное девушками-нефилологами.

Началась и публикация материалов сбалансированной части ЗКРЯ (см. *Русская спонтанная речь* 2008, 2010а,б), что делает их доступными широкому кругу исследователей самого разного ранга.

Литература

1. *Асиновский А. С., Богданова Н. В., Русакова М. В., Степанова С. Б., Шерстинова Т. Ю.* Звуковой корпус русского языка повседневного общения «Один речевой день»: концепция и состояние формирования // Компьютерная лингвистика и интеллектуальные технологии. Выпуск 7 (14). По материалам ежегодной международной конференции «Диалог» (2008) / Гл. ред. А. Е. Кибрик. М., 2008 С. 488–494.
2. *Асиновский А. С., Богданова Н. В., Русакова М. В., Рыко А. И., Степанова С. Б., Шерстинова Т. Ю.* Звуковой корпус как способ мониторинга и фиксации разных форм естественного языка // Компьютерная лингвистика и интеллектуальные технологии. Выпуск 8 (15). По материалам ежегодной международной конференции «Диалог» (2009) / Гл. ред. А. Е. Кибрик. М., 2009. С. 38–44.
3. *Богданова Н. В.* Уровень речевой компетенции как реальная социальная характеристика говорящего, определяющая его речь // Материалы XXXVIII международной филологической конференции. Выпуск 22. Полевая лингвистика. Интегральное моделирование звуковой формы естественных языков. 16–20 марта 2009 года / Отв. ред. А. С. Асиновский, науч. ред. Н. В. Богданова. СПб., 2010а. С. 29–40.
4. *Богданова Н. В.* Вставные конструкции в звучащем спонтанном монологе (к проблеме построения грамматики русской речи) // Вопросы культуры речи. М., 2010б (в печати).
5. *Богданова Н. В., Бродт И. С., Куканова В. В., Павлова О. В., Сапунова Е. М., Филиппова Н. С.* О «корпусе» текстов живой речи: принципы формирования и возможности описания // Компьютерная лингвистика и интеллектуальные технологии. Выпуск 7 (14). По материалам ежегодной международной конференции «Диалог» (2008) / Гл. ред. А. Е. Кибрик. М., 2008. С. 57–61.
6. *Бродт И. С.* Спонтанный монолог в лингвистическом и социолингвистическом аспектах (на материале текстов разного типа). Дис. … канд. филол. наук. СПб., 2007 (машинопись).
7. *Иванова О. А.* К характеристике внутриязыкового контакта между литературной и профессиональной речью носителя русского языка // Материалы XXXVII международной филологической конференции. Выпуск 21. Полевая лингвистика. Интегральное моделирование звуковой формы естественных языков. 10–15 марта 2008 года / Отв. ред. А. С. Асиновский, науч. ред. Н. В. Богданова. СПб., 2008. С. 25–35.
8. *Иванова О. А.* Стилевая «разноголосица» в бытовой спонтанной русской речи // Вестник Санкт-Петербургского университета. Филология. Востоковедение. Журналистика. Серия 9. СПб., 2010 (в печати).
9. *Куканова В. В.* Лингвистический анализ репродуцированных текстов (на материале звукового корпуса русской речи юристов). Дис. … канд. филол. наук. СПб., 2009 (машинопись).
10. *Русская спонтанная речь.* Свободные монологи-рассказы на заданную тему. Тексты. Лексические материалы / Сост. В. В. Куканова / Отв. ред. и автор предисловия Н. В. Богданова. СПб., 2008.
11. *Русская спонтанная речь.* Монологи-репродуктивы. Тексты. Лексические материалы / Сост. В. В. Куканова / Отв. ред. и автор предисловия Н. В. Богданова. СПб., 2010а (в печати).
12. *Русская спонтанная речь.* Монологи-описания. Тексты. Лексические материалы / Сост. В. В. Куканова / Отв. ред. и автор предисловия Н. В. Богданова. СПб., 2010б (в печати)/
13. *Сапунова Е. М.* Неподготовленное чтение как вид речевой деятельности и тип устного спонтанного монолога (на материале русского языка) Дис. … канд. филол. наук. СПб., 2009 (машинопись).
14. *Степихов А. А.* Соотношение синтаксического и интонационного членения в спонтанном монологе. Дис. … канд. филол. наук. СПб., 2005 (машинопись).
15. *Филиппова Н. С.* Принципы построения устного описательного дискурса (на материале русской спонтанной речи). Дис. … канд. филол. наук. СПб., 2010 (машинопись).

Звуковой корпус русского языка «один речевой день»: пути пополнения и первые результаты исследования[1]

The ord speech corpus of russian everyday communication: ways of replenishment and first results of analysis

Богданова Н. В. (nvbogdanova_2005@mail.ru),
Асиновский А. С. (a.s.asinovsky@gmail.com),
Маркасова Е. В. (markasovaelena@yandex.ru),
Степанова С. Б. (stsvet_2002@mail.ru),
Супрунова А. В. (nsuprunova@mail.ru),
Шерстинова Т. Ю. (sherstinova@gmail.com)

Факультет филологии и искусств Санкт-Петербургского государственного университета, Санкт-Петербург, Россия

В докладе представлены некоторые конкретные результаты анализа материалов звукового корпуса «Один речевой день», а также пути его пополнения и дальнейшего развития.

Корпус «Один речевой день» (ОРД) представляет собой один из блоков Звукового корпуса русского языка (ЗКРЯ), работа над которым ведется на факультете филологии и искусств СПбГУ. Принцип, положенный в основу организации этой части корпуса, условно можно назвать «*принципом невода*»: забрасываем широкую сеть в средуносителей языка, вытягиваем все, что в нее попало, и делаем это объектом многоуровневого анализа. Предложенная в определении принципа метафора легко поддерживается высказываниями авторитетных лингвистов, которые сравнивают народный разговорный язык (т. е. речь) с широкой и мощной рекой, ледяной покров которой — это письменность (см.: Вандриес 1937: 253; Bally 1952: 13). Еще Л. В. Щерба писал о необходимости «эмансипации от письменного языка и обращения к живой речи» (Щерба 1974: 144). Именно в целях такого обращения к живой речи в ее свободном функционировании, не ограниченном ни лабораторными условиями записи, ни конкретными речевыми заданиями для информантов, и была применена методика 24-часовой записи (подробнее о корпусе ОРД см.: Степанова и др. 2008; Богданова и др. 2009). *Принцип невода* позволяет увидеть реальную, естественную, а не искусственно созданную в лабораторных условиях, жизнь, отраженную в речи.

Работа над этой частью корпуса продолжается в настоящее время по двум направлениям: пополнение его новыми записями и анализ существующего материала. Целью настоящего сообщения является как раз представление новых поступлений в корпус, первые конкретные результаты лингвистического анализа материала на разных уровнях и определение перспектив дальнейших исследований.

Новые записи «Одного речевого дня» были осуществлены с целью сбалансировать гендерный состав информантов. Так, первый «заброшенный невод» принес вполне ожидаемый в этом отношении «улов»: из 30 информантов — участников пилотного эксперимента — оказалось 11 мужчин и 19 женщины, что вполне отражает гендерную ситуацию в нашем обществе, однако снижает возможности сопоставительного описания речи мужчин и женщин в сходных

[1] Исследование выполнено при поддержке гранта РГНФ «Разработка информационной среды для мониторинга устной русской речи» (09-04-12115в).

коммуникативных ситуациях. Для выравнивания (балансировки) этого состава были записаны еще 9 информантов-мужчин, в основном старшего возраста, и 1 информант-женщина. Некоторое отступление от выдвинутого принципа невода сполна окупается тем, что во всех остальных отношениях он остается ведущим при формировании ОРД. В наибольшей степени принцип невода работает, когда речь идет не о самих информантах, а об их коммуникантах, число которых достигает в настоящее время 600 и состав которых — не только гендерный, но также социологический и психологический — не поддается никакой балансировке. В результате в зоне внимания исследователей, авторов и разработчиков этого проекта, оказываются и записи речевого материала (также далекие от всякой балансировки — тематической, ситуативной и т. п.), и все характеристики всех информантов и их коммуникантов, что открывает почти безграничные возможности для лингвистического (в том числе социо- и психолингвистического) анализа современной русской речи.

Другой путь пополнения ОРД и в целом развития исследований в этом направлении — это некоторое пересечение материала двух блоков ЗКРЯ: данного корпуса повседневной речи носителей языка и сбалансированной части корпуса, устроенной, в отличие от первой, по «принципу ковчега» (см. о ней подробнее: Богданова и др. 2008, а также статью Н. В. Богдановой в настоящем сборнике). Так, одним из новых участников эксперимента стал информант (девушка-студентка, типичный экстраверт), речь которого до этого записывалась многократно в течение трех лет по одной и той же лингвистической программе (чтение, пересказ, описание изображения, свободный рассказ на заданную тему). Каждый речевой сценарий повторялся 4 раза, с интервалом в 2 месяца, что позволило выявить особенности не конкретного речевого акта, а речи данной языковой личности в целом (создать речевой портрет человека). Даже с учетом того факта, что и на этот раз информант знал о проводящейся записи, его речь следует признать значительно более естественной по сравнению с речью в микрофон, в присутствии исследователя. Таким образом появилась возможность сравнить речь одного говорящего в разных условиях записи: В перспективах этого направления исследования — повторные (как минимум, еще одна, через год) записи речевого дня данного информанта, для выявления всех особенностей его речевого поведения и уточнения его речевого портрета.

Любопытным представляется и опыт записи ОРД от информанта-иностранца, живущего в данное время в России. Таким информантом стала студентка из Швейцарии, приехавшая учиться в СПбГУ. Первая запись ее ОРД была осуществлена в ноябре 2009 г., в первые месяцы пребывания в стране изучаемого языка. Эта запись продемонстрировала не только определенный уровень владения информантом русским языком, а также смену языков (русский, немецкий, французский) в различных коммуникативных ситуациях, но и специфику обращенной к нему речи его коммуникантов, носителей русского языка. Повторные записи ОРД от того же информанта, которые планируются на ноябрь 2010 и 2011 гг., позволят увидеть не только динамику овладения иностранцем русской речью (на всех уровнях), но и изменение особенностей речи его коммуникантов, которые невольно реагируют на степень интерферированности речи собеседника.

В целом корпус ОРД характеризуется в настоящее время такими количественными показателями: 320 часов звучания, полученные от 40 информантов (20 мужчин и 20 женщин). Звукозаписи переформатированы, убраны длительные (больше 5 минут) шумовые фрагменты, не содержащие речи. Звукозаписи разрезаны на коммуникативные эпизоды по принципу общих условий коммуникации и качества звукозаписи (см.: Asinovsky et al. 2009). В результате было получено 994 файла-эпизода общей продолжительностью 268 часов. Подготовлены методики многоуровневого аннотирования данных на лингвистическом и паралингвистическом уровнях (см.: Шерстинова и др. 2009). Общая структура корпуса «Один речевой день» описана в работе: Sherstinova 2009a.

Осуществлена расшифровка 34 часов звукозаписей для 40 информантов и их многоуровневое аннотирование в профессиональной программе ELAN. В результате расшифровки получены тексты общим объемом в 244 075 словоупотреблений на уровне Frase (реплики говорящих), которые относятся к 125 эпизодам и соответствуют 33,87 часам непрерывной звучащей речи. Для 20 информантов-мужчин получено 106 109 словоупотреблений на уровне реплик, что соответствует 15,10 часам непрерывной речи, для 20 информантов-женщин — 137 966 словоупотреблений, что соответствует 18,77 часам.

Аннотирование осуществлялось по следующим уровням:

- Phrase (реплики говорящих),
- Speaker (код говорящего),
- Events (невербальные аудиособытия),
- Voice (качество голоса говорящего),
- PhonetCom (фонетический комментарий),
- PhraseComment (фразовый комментарий),
- Notes (общий комментарий),
- Episode (мини-эпизод речевой коммуникации).

Реплики говорящих содержат синтагматическое и фразовое членение.

По материалам расшифрованных записей 40 информантов был получен частотный словник всех использованных в речи словоформ и выполнен первичный анализ этих данных (Sherstinova 2009b). Полученные количественные данные о частоте употребления каждой словоформы были перенесены в программу MS Excel и пропущены через сортировку от максимального к минимальному показателю упо-

требительности. Единицей описания в настоящем исследовании признается не только графическое слово («от пробела до пробела»), с которым работают многие морфологические анализаторы и программы по созданию конкорданса, но и знаки сегментирования, а также символы, обозначающие паралингвистические явления, сопровождающие естественную речь (<смех>, <кашель>, <вздох> и под.).

Параллельно с созданием корпуса начаты исследования фонетики и грамматики русской повседневной речи, а также особенностей повседневной коммуникации (см.: Богданова 2008; Королева 2009; Маркасова 2009а; Маркасова, Корякова 2008; Асиновский 2010; Королева, Журавлева 2010; Рыко, Степанова 2010).

Расшифрованный материал корпуса включает более чем 32 000 фраз. Подавляющее большинство высказываний в материале исследования оказались утвердительными (23 494), восклицательных и вопросительных высказываний встретилось около девяти тысяч (см. табл. 1. и рис. 1).

Таблица 1. Абсолютные данные сегментирования в материале ОРД

Значение символа	**Знак сегментирования**	**Абсолютное количество**
Знак членения на повествовательные высказывания	//	23 494
Знаки для обозначения восклицательных и вопросительных высказываний	?	6627
	!	2080
ИТОГО		**32 201**

Как хорошо видно на диаграмме (см. рис. 1), говорящий в течение своего речевого дня гораздо чаще нечто сообщает собеседникам, лишь четвертая часть всех высказываний в материале корпуса — вопросительные и восклицательные конструкции.

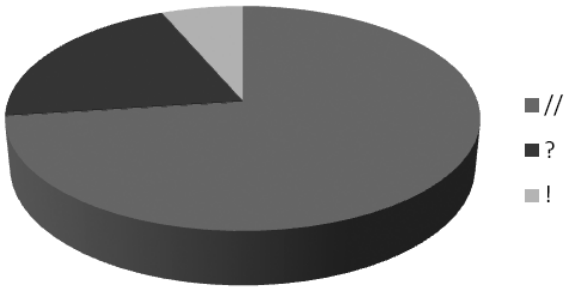


Рисунок 1. Соотношение различного рода высказываний в материале ОРД

Начато сегментирование речевого материала на лексическом уровне. Осуществлена сегментация 14 884 слов для 5 информантов (S01, S19, S24, S35, S37) (147 минут звучания).

Получена выборочная реальная транскрипция для 1000 словоизменительных морфем из речи 4 информантов. Морфемное аннотирование включает три уровня — орфографическую запись морфемы, её функциональный тип и реальную фонетическую транскрипцию.

Что касается конкретных результатов исследования материалов этой части ЗКРЯ, то они связаны с разными уровнями анализа и частично представлены в настоящем сборнике (см. доклад Е. В. Маркасовой и С. А. Воробьёвой о функционировании в корпусе ОРД дискурсивного слова *конечно*, а также доклад А. С. Асиновского и др. «Звуковая реальность словоизменительных аффиксов (по данным Звукового корпуса русского языка)»).

Весьма перспективным представляется анализ лексики на материале ОРД, сравнение её частотности и контекстуальных значений с тем, что представлено в Национальном корпусе русского языка (НКРЯ) и словарях, построенных на материале письменной речи. Так, анализ дискурсивной лексики в ОРД позволил констатировать постепенный «уход» из повседневной речи таких вводных слов, выражающих (по классическим их определениям) оценку говорящим степени достоверности сообщаемого, как *несомненно, бесспорно, вне всякого сомнения*, а встретившееся *безусловно* можно рассматривать, скорее, как маркер манипулятивного речевого поведения личности, имеющей иллокутивную установку доминирования (см. Маркасова 2009а).

Аналогичная ситуация с «маркерами искренности»: такие конструкции, как *говоря по совести, чего скрывать, положа руку на сердце* и под., не встретились в корпусе ОРД ни разу. Конструкции же, достаточно частотные в повседневной речи — *честно говоря* и ее вариант *(если) честно*, — в языке повседневного общения используются не в прямом значении, то есть не в качестве маркера искренности в речевом акте признания, а в качестве контактоустанавливающего элемента, вносящего в высказывание семантику солидаризации или интимизации общения (см. примеры 1, 2, 3), или как заполнитель паузы хезитации (примеры 4, 5) (см. Маркасова 2009б):

(7) *у меня вот* ***честно говоря*** */ желание / только <…> / пойти (…) влить / в себя что-нибудь … / чтоб повеселело*;

(8) *я тоже* ***честно говоря*** *с удовольствием кофе бы попил*;

(9) *по-моему попьем и не умрем //* ***честно говоря*** */ я дома тут пристрастился тоже Нурю пить*;

(10) *да ну не знаю /* ***честно говоря*** *не помню*;

(11) *я* ***честно говоря*** *// во-первых ну я могу конечно показать вам эти анкеты.*

Проведенный анализ фразеологических оборотов (ФО), используемых в повседневной речи носителей русского языка, также выявил ряд интересных особенностей их функционирования (см. подробнее: Супрунова 2010). Количество употребляемых фразеологизмов, как показало исследование, напрямую связано с их лингвистическими характеристиками. Можно говорить о достаточно высокой частоте употребления ФО в живой речи, однако это относится лишь к определенным их типам.

Так, книжные фразеологизмы в спонтанной речи (СР) почти не употребляются (2 % от общего количества ФО в расшифровках ОРД) (*завершающий аккорд; притча во языцех; пожинать плоды; камень преткновения*), тогда как разговорных и просторечных ФО достаточно много (68 %) (*челюсть отвисла; крышняк едет; вынь да положь; поджилки трясутся; дать на лапу; галопом по Европам; ни в одном глазу*).

В живой речи практически отсутствуют фразеологические сочетания (0,3 %) (*В кромешной тьме*) и мало фразеологических сращений (14 %) (*губа не дура; дойти до ручки; ни в одном глазу; плести всякую ахинею*), при большом количестве фразеологических единств (35 %) (*стоять над душой; раз плюнуть; ставить на карту*) и фразеологических выражений (51 %) (*в крайнем случае; будь другом; время от времени; изо всех сил; посреди белого дня).*

Носители языка в целом предпочитают с помощью ФО характеризовать в своей речи действия, поступки человека (17 %), а также выражать эмоции (14 %), что обусловливает преобладание в СР фразеологических оборотов следующих лексико-грамматических разрядов:

- ☐ адвербиальные (27 %) (*задним числом; всеми силами души; кровь из носу*),
- ☐ глагольные (18 %) (*делать крюк; дать сдачи; убить время; лезть в дебри*);
- ☐ междометные (21 %) (*не дай бог; с ума сойти!; ничего себе!; слава богу; мать моя женщина*).

Что касается эмоционально-оценочной коннотации, то в живой речи используется небольшое количество ФО с положительной оценкой (12 %) (*первый парень на деревне*; *голубая кровь*; *не из робкого десятка*) и очень мало ФО с эмоционально-экспрессивной окраской (ЭЭО):

- ☐ осуждение (0,6 %) (*на иглу сажали; денег некуда девать*),
- ☐ восхищение (1,2 %) (*владеет словом; как игрушка*).

Носители языка стремятся быть достаточно сдержанными в своей речи, поэтому в ней доминируют фразеологические обороты с нейтральной коннотацией (57 %), без ЭЭО (70 %) *(приходить в голову; асфальтовая болезнь; изо всех сил; один к одному; держать на замке*).

Тем не менее, в речи информантов встретилось большое количество ФО с отрицательной коннотацией (28 %) (*понты раскинуть; горло рвать; потерять себя; глухая тетеря; дошла до ручки; на ладан дышать*). Из всех фразеологических оборотов с ЭЭО количественно преобладают единицы с не столь резкими, категоричными, в сравнении с другими, типами окраски:

- ☐ шутливость (10 %) (*на халяву и уксус сладкий; трудовая мозоль; глухая тетеря; два брата-акробата*),
- ☐ неодобрение (5 %) (*потерять себя; убить время; сидеть на шее*).

Не последнее место среди возможностей использования материалов ЗКРЯ занимает *создание лексикографического описания бытовой спонтанной звучащей речи*, что является еще одной перспективной задачей настоящего проекта. Думается, что такое описание может вестись, как минимум, в трех направлениях.

1. *Словарь русской бытовой разговорной речи*. Возможности для создания такого словаря дает *электронная картотека E-Kar*, одно из программных средств, использованных при построении Звукового корпуса. Эта программа автоматически создает конкорданс по выбранным текстам и позволяет решать многообразные задачи классификации и описания языковых единиц. В частности, она позволяет собрать по тексту все словоформы, в нем встречающиеся, посчитать их частоты, предъявить для этих словоформ любое лингвистическое расширение по тексту или группе текстов, имеющихся в электронной коллекции. Результаты автоматической работы программы (конкорданса) и последующей работы эксперта дают возможность новых содержательных форм интерпретации лингвистического материала. Необходимость и целесообразность создания такого словаря можно проиллюстрировать простым примером. В существующих к настоящему времени расшифровках материалов ЗКРЯ обнаружилось всего два употребления слова *хлеб*, оба — в значении ‘кусок хлеба’ в контекстах одинакового типа: *намазать хлеб маслом*. Проверка по материалам МАС показала, что подобное — весьма и весьма распространенное — значение данного (очень частотного) слова в Академическом словаре попросту отсутствует. Электронная версия словаря русской бытовой разговорной речи позволит не только увидеть все новые значения привычных слов и получить их контексты, но и — при необходимости — услышать их звучание в реальной речи, что обеспечивается другими программными средствами, использованными при создании ЗКРЯ.
2. *Словарь контекстных экспрессем русской разговорной речи*. Уже первые шаги в реализации этого проекта убеждают в том, насколько он может быть интересен и полезен самым разным специалистам по изучению русской речи.

Вот, например, как выглядят фрагменты корпуса, размеченные с лексикографической точки зрения:

- *она привозила с собой этого немца / который* () *вот* ***явный немец*** */ а* ***косил под русского Ваньку*** *// ему первые два дня было запрещено говорить // потом* (…) *он… под дурачка косил* ? *он был* ***первым парнем на деревне*** *// там к сему* (…) */ к нему все девки* (…) *пока не проговорился к нему все девки приставали // пока не проговорился / да / по-немецки / пока молчал // угу // это я на колонке угостил (э-э) немца // салом // выпьет водки // он долго не хотел / а потом как* ***подсел на*** *это* ***сало*** */ короче // и давай* ***один бутерброд за другим наворачивать*** *// просто за друг…* (…) *перед другими людьми уже неудобно было // так* ***попёрло его***;
- *Дора /* ***типа*** *привет / у меня к тебе* ***шкурная тема*** *Коля / и какая у тебя* ***шкурная тема*** *ко мне? ну* ***типа*** */ причем разговаривает в каком-то таком стиле //* ***красавица с накладными мозгами***;
- *они (э) дела… стараются делать для всех / то есть / невзирая на личности / да ? угу просто / что видит человека / вот* ***из него*** *можно* ***доить денежку*** *// вот например торговый представитель / определённый //* ***с него*** *можно* ***тянуть денежку*** */ определённую // вот он сейчас… / его запустить угу /* ***выжать из него все соки*** */ получить с него максимум /* (*а…) прибыли / и его можно выкинуть*.

Видно, как много материала для лексикографического описания экспрессивной лексики содержат материалы ЗКРЯ.

3. Наконец, третье направление лексикографического описания материалов ЗКРЯ может быть связано с *корпусом редуцированных форм* (РФ) русской речи (*рупь, дяинька, пийсят, прально, тода, то ись, кагорится, собссно гря* и под.). В известном смысле это направление пересекается со словарем русской бытовой разговорной речи (см. выше п. 1), поскольку электронная версия такого словаря предполагает возможность прослушать любое слово в любом контексте:

- *просто не горит // а чек ты в сумку не* ***хошь*** *положить?*
- ***здрасть*** */ отдел кадров уже закрыт? ага /* ***сёння*** *же пятница*;
- *а-а вот мы / живем / на Смолячкова /* ***ничо*** *не знаем*.
- *мы как раз здесь побудем / за час уже выболтается всё // а он же там* ***буит****… это*;
- *м-м хорошо / давай // ты во* ***скоко*** *будешь дома?*

Наличие словаря РФ, словник которого построен по алфавиту исходных словарных форм, позволит целенаправленно искать ту или иную единицу в материале корпуса и получать, например, статистику не только употребительности того или иного слова или словосочетания, но и представленности в нашей повседневной речи всех реальных форм языковых единиц. Образец словарной статьи подобного словаря см. в табл. 2.

Таблица 2. Структура словарной статьи словаря редуцированных форм русской речи

Лексема	Частеречное значение	Варианты написания	Примеры из НКРЯ и др. письменных источников	Частотность (если возможно)	Варианты произношения	Примеры из ЗКРЯ	Частотность	Специфика функционирования	Специфика значения	Примечания

Такой словарь может не только быть полезен специалистам по устной спонтанной речи, но и стать неоценимым помощником преподавателя русского языка в любой аудитории, ориентированной на знакомство с современной русской речью (см. о нем подробнее: Богданова, Пальшина 2010).

По мере сбора и обработки материалы ЗКРЯ традиционно передаются в устный подкорпус Национального корпуса русского языка («Из материалов корпуса “Один речевой день”, подготовленного группой А. С. Асиновского»), что делает их доступными для самого широкого круга пользователей.

Литература

1. *Асиновский А. С.* Как звучат русские флексии // Материалы XXXVIII международной филологической конференции. Выпуск 22. Полевая лингвистика. Интегральное моделирование звуковой формы естественных языков. 16–20 марта 2009 г. Санкт-Петербург / Отв. ред. А. С. Асиновский, науч. ред. Н. В. Богданова. СПб., 2010. С. 4–28.
2. *Богданова Н. В.* Аллегровые формы русской речи: от произносительной редукции к письменной фиксации и лексикализации в языке // Материалы XXXVII международной филологической конференции. Фонетика. 10–15 марта 2008 года / Отв. ред. Н. Д. Светозарова. СПб., 2008. С. 31–44.
3. *Богданова Н. В., Асиновский А. С., Русакова М. В., Рыко А. И., Степанова С. Б., Шерстинова Т. Ю.* Звуковой корпус как способ монито-

ринга и фиксации разных форм естественного языка // Компьютерная лингвистика и интеллектуальные технологии. Выпуск 8 (15). По материалам ежегодной международной конференции «Диалог» (2009) / Гл. ред. А. Е. Кибрик. М., 2009. С. 38–44.

4. *Богданова Н. В., Бродт И. С., Куканова В. В., Павлова О. В., Сапунова Е. М., Филиппова Н. С.* О «корпусе» текстов живой речи: принципы формирования и возможности описания // Компьютерная лингвистика и интеллектуальные технологии. Выпуск 7 (14). По материалам ежегодной международной конференции «Диалог» (2008) / Гл. ред. А. Е. Кибрик. М., 2008. С. 57–61.
5. *Богданова Н. В., Пальшина Д. А.* Редуцированные формы русской речи (опыт лексикографического описания) // Материалы ежегодной Всероссийской научной конференции «Слово. Словарь. Словесность: Текст словаря и контекст лексикографии». Санкт-Петербург, 11–13 ноября 2009 г. СПб., 2010 (в печати).
6. *Вандриес Ж. Язык. М., 1937.*
7. *Королева И. В.* Коммуникативное взаимодействие: смена стратегии в зависимости от психосоциальных особенностей говорящих (на материале Звукового корпуса русского языка «Один речевой день») // Речевая коммуникация в современной России: Материалы I международной научной конференции (Омск, 27–29 апреля 2009 г.) / Под ред. О. С. Иссерс, Н. А. Кузьминой. Омск, 2009. С. 155–162.
8. *Королева И. В., Журавлева А. А.* Личностные характеристики говорящих и их влияние на коммуникативное взаимодействие (на материале Звукового корпуса русского языка «Один речевой день») // Материалы XXXVIII международной филологической конференции. Выпуск 22. Полевая лингвистика. Интегральное моделирование звуковой формы естественных языков. 16–20 марта 2009 г. Санкт-Петербург / Отв. ред. А. С. Асиновский, науч. ред. Н. В. Богданова. СПб., 2010. С. 76–89.
9. *Маркасова Е.* Слово-резонер «безусловно» // Auspicia № 2, 2009a.
10. *Маркасова Е. В.* Маркеры искренности в языке повседневности (признаться сказать, говоря по совести, по чести говоря, честно говоря) // Rossica Olomucensia 2009б. Vol. XLVIII. № 2. S. 149–155.
11. *Маркасова Е. В., Корякова Е. В.* Речевая агрессия и диагностика личностной акцентуации (по данным корпуса «Один речевой день») // Материалы XXXVII международной филологической конференции. Выпуск 21. Полевая лингвистика. Интегральное моделирование звуковой формы естественных языков. 10–15 марта 2008 года / Отв. ред. А. С. Асиновский, науч. ред. Н. В. Богданова. СПб., 2008. С. 62–70.
12. *Рыко А. И., Степанова С. Б.* Индивидуальные стратегии членения спонтанной речи на предложения (по результатам перцептивного эксперимента) // Материалы XXXVIII международной филологической конференции. Выпуск 22. Полевая лингвистика. Интегральное моделирование звуковой формы естественных языков. 16–20 марта 2009 г. Санкт-Петербург / Отв. ред. А. С. Асиновский, науч. ред. Н. В. Богданова. СПб., 2010. С. 128–137.
13. *Степанова С. Б., Асиновский А. С., Богданова Н. В., Русакова М. В., Шерстинова Т. Ю.* Звуковой корпус русского языка повседневного общения «Один речевой день»: концепция и состояние формирования // Компьютерная лингвистика и интеллектуальные технологии. Выпуск 7 (14). По материалам ежегодной международной конференции «Диалог» (2008) / Гл. ред. А. Е. Кибрик. М., 2008. С. 488–494.
14. *Супрунова А. В.* Закономерности употребления фразеологических оборотов в устной речи (на материале Звукового корпуса русского языка) // Материалы XXXVIII международной филологической конференции. Выпуск 22. Полевая лингвистика. Интегральное моделирование звуковой формы естественных языков. 16–20 марта 2009 г. Санкт-Петербург / Отв. ред. А. С. Асиновский, науч. ред. Н. В. Богданова. СПб., 2010. С. 147–155.
15. *Шерстинова Т. Ю., Степанова С. Б., Рыко А. И.* Система аннотирования в звуковом корпусе русского языка «Один речевой день» // Материалы XXXVIII международной филологической конференции. Формальные методы анализа русской речи. 16–20 марта 2009 г. Санкт-Петербург. СПб., 2009 (в печати).
16. *Щерба Л. В.* О разных стилях произношения и об идеальном фонетическом составе слова // Л. В. Щерба. Языковая система и речевая деятельность. Л., 1974. С. 141–146.
17. *Asinovsky A., Bogdanova N., Rusakova M., Stepanova S., Ryk A., Sherstinova S.* The ORD Speech Corpus of Russian Everyday Communication «One Speaker's Day»: Creation Principles and Annotation // LNCS/LNAI series. «Text, Speech and Dialogue» TSD-2009. Springer-Verlag, Berlin, Heidelberg. 2009. Pp. 250–257.
18. *Bally Ch.* Le langage et la vie. Genève-Lille, 1952.
19. *Sherstinova T.* The Structure of the ORD Speech Corpus of Russian Everyday Communication // «Text, Speech and Dialogue» TSD-2009. LNAI 5729. Springer-Verlag, Berlin, Heidelberg. 2009a. Pp. 258–265.
20. *Sherstinova T.* Quantitative data processing in the ORD speech corpus of Russian everyday communication // Proc. of International Quantitative Linguistics Conference QUALICO 2009. Graz, Austria. 2009b. Pp. 56–57.

О валентных свойствах одного широкого класса существительных[1]

On valency properties of a wide class of nouns

Богуславский И. М. (bogus@iitp.ru), **Иомдин Л. Л.** (iomdin@iitp.ru)
Институт проблем передачи информации им. А. А. Харкевича РАН

Рассматривается валентная структура нескольких лексикографических типов русских предикатных существительных, конституирующей особенностью которых является тот факт, что один из актантов всякого такого существительного может именоваться самим этим словом. Например, во фразе *Расстояние в 650 километров этот поезд преодолевает за 4 часа* выражение *650 километров* заполняет валентность слова *расстояние,* но *650 километров* — это и есть *расстояние.*

1. Виды заполнения валентностей слов

Мы будем исходить из того, что в синтаксической структуре предложения актанты могут занимать разные позиции относительно предикатного слова. Существуют три основных типа заполнения семантической валентности слова: активный, пассивный и разрывный [4].

Канонический способ заполнения некоторой валентности предикатного слова *L* состоит в том, что актант *A* синтаксически подчиняется слову *L,* например, *мальчик* (*A*) *спит* (*L*). Такое заполнение валентности мы называем **активным**.

Наряду с этим весьма массовыми являются ситуации, когда актант *A* синтаксически подчиняет *L*; такое заполнение валентности называется **пассивным**. Типичным примером здесь является заполнение валентностей у прилагательных; скажем, в выражении *странный* (*L*) *человек* (*A*) слово *человек* заполняет семантическую валентность прилагательного *странный*, но синтаксически подчиняет его.

Наконец, между самим словом *L* и некоторым его актантом может вообще не быть непосредственной синтаксической связи. Такое заполнение валентности мы будем называть **разрывным**. Например, в выражениях *любимый* (*L*) *режиссёр Ватикана* (*A*) или *мой* (*A*) *любимый* (*L*) *режиссёр*: слова *Ватикана* и *мой* реализуют субъектную валентность прилагательного *любимый,* но синтаксически не подчиняют это прилагательное и не подчиняются ему, а зависят от третьего слова *режиссёр* (от которого «параллельно» синтаксически зависит и данное прилагательное).

Весьма часты также случаи, когда при некотором слове разные валентности заполняются разными способами. Таковы, например, валентности кванторных слов. В частности, в предложении *Все* (*L*) *дети* (*A1*) *капризничают* (*A2*) один актант слова *все* — *A1* — присоединяется к нему пассивно, а другой — *A2* — разрывно.

Интересно, что предикатных слов, все валентности которых заполняются только активным образом, не так уж много: эти слова явно составляют меньшинство. В самом деле, даже глаголы синтаксически подчиняют все свои актанты только тогда, когда они стоят в личной форме: в противном случае актанты, соответствующие глагольному подлежащему (т. е. представляющие субъектную или, в случае страдательного залога, объектную валентность глагола) реализуют эти валентности пассивным или разрывным образом, ср. *Мальчик* (*A*) *спит* (*L*), но *спящий* (*L*) *мальчик* (*A*) или *Мальчик* (*A*) *хотел спать* (*L*).

Разрывное заполнение валентностей изучено относительно слабо и представляет особый интерес. Спектр этого явления довольно широк — от отдельных слов типа *любимый* и узких групп типа кванторных слов до лексически неограниченных конструкций, включающих инфинитив глагола. Ниже мы обсудим один весьма широкий класс слов, для которых разрывное заполнение одной из валентностей является стандартным. В него входят такие разные слова, как *учитель, отец, однополчанин, цель, привычка, скорость, имя, фамилия, копия, оригинал, версия, вариант* и многие другие. Эти слова обладают важным общим свойством, которое позволяет объединить их в единый класс.

[1] Авторы выражают благодарность Российскому фонду фундаментальных исследований (гранты №№ 08-06-00344 и 08-06-00373) за частичную финансовую поддержку данной работы.

2. Актантные автодериваты

В соответствии с концепцией Московской семантической школы предикатные слова, т. е. слова, имеющие валентности, должны толковаться в составе выражения, в котором естественным образом были бы отражены все актанты толкуемого слова. Так, входом толкования глагола *обжечь* должно служить не просто само это слово, а выражение

(12) *X обжег Y Z-ом*

(например, *Маша обожгла руку сковородкой*), в котором все валентности глагола представлены переменными. На начальных этапах изучения валентностей внимание было сосредоточено на прототипических предикатных словах — глаголах. Поэтому толкуемое выражение всегда имело вид целого предложения и получило название сентенциальной формы. Впрочем, вскоре выяснилось, что толкуемая форма не всегда должна быть сентенциальной. Так, существительное *ожог* имеет тот же набор валентностей, что и глагол *обжечь*, и все они отлично выразимы в составе именной группы:

(13) *ожог у X-а на Y-е от Z-а* (*ожог на руке у Маши от сковородки*).

Обратимся к словам типа *учитель, ученик, врач, пациент, покупатель, продавец* и т. п., которые относятся к классу актантных дериватов S_i. Они подробно описаны в «Лексической семантике» Ю. Д. Апресяна [1], с. 164–168. Эти слова относятся к той же ситуации, что и глаголы *учить, лечить* или *покупать*, но служат обозначением не самой этой ситуации, как глаголы, а некоторого ее актанта. Дериваты типа S_i имеют те же семантические валентности, что и исходный глагол, за исключением того, что *i*-я валентность у них принципиально не выражается зависимым от них словом. Поэтому входом толкования для таких слов служит связочное предложение, в котором актант, не выразимый непосредственно при толкуемом слове, присоединяется к нему через связку: *X (есть) учитель Y-а* и т. д.

Разумеется, далеко не все связочные предложения обладают тем свойством, что одно из существительных, зависящих от связки, заполняет валентность другого. Ср., например, предложения *Береза — дерево, Это дерево — береза*, в которых ни одно из существительных не является актантом другого.

Тем не менее слов, способных присоединять один из своих актантов через связку, довольно много, и далеко не все из них являются синтаксическими дериватами от какого-то другого слова.

Один из классов подобных слов образуют существительные, обозначающие разнообразные отношения между людьми, такие, как *отец, мать, сестра, брат, коллега, однополчанин, земляк, единомышленник, конкурент, противник* и многие другие. В качестве выражения, служащего входом толкования для таких слов, также естественно использовать связочное предложение, например,

(14) *X (есть) однополчанин Y-а* = ‘человек X служит или служил в том же полку, что и Y’.

Для некоторых из слов — имен отношений существуют соответствующие глаголы, например, *друг — дружить, враг — враждовать*. Однако для большинства слов этого класса подобных глаголов не существует, и поэтому эти слова не могут считаться синтаксическими дериватами типа S_i.

Следующую интересную группу образуют имена абстрактных соотношений и свойств типа *причина, цель, результат, следствие, идея, замысел, привычка, обычай, новость, тайна* и т. п., например:

(15) *Съедать* (*A*) *на завтрак тарелку овсяной каши было его давней привычкой* (*L*).

Добавим к этим лексикографическим типам существительных группу параметров типа *длина, ширина, грузоподъемность, температура, национальность, гражданство, длительность, время, координаты, место, местонахождение* и т. п., ср.

(16) *Длина* (*L*) *доски — три метра* (*A*).

Некоторые параметрические существительные имеют соответствующий глагол и поэтому могут считаться принадлежащими к классу S_i от таких глаголов, например, *вес — весить*. Однако для большинства слов этой группы наличие такого глагола нетипично.

Тот факт, что все эти слова способны присоединять один из своих актантов через связочный глагол, по существу означает, что все они способны **называть** один из своих актантов. Поясним, что мы имеем в виду.

Вернемся к слову *ожог*, которое мы упоминали в примере (2). Ситуация, обозначаемая этим словом, включает трех участников — 1) живое существо, 2) его часть тела, которая вошла в контакт с объектом аномально высокой температуры или химически активным веществом, и 3) сам этот объект или вещество: *ожог на руке у Маши от сковородки*. Существенно, что ни один из этих актантов — ни Маша, ни ее рука, ни сковородка — не может быть назван ожогом!

В случае существительных перечисленных выше классов дело обстоит иначе. Если один из актантов слова *учитель* можно назвать учителем (*Иван — учитель Петра*), то и один из актантов *однополчанина* можно назвать однополчанином (*Иван — мой однополчанин*), один из актантов *цели* можно назвать целью (*Их цель — построить вечный*

двигатель), а один из актантов слова *длина* можно назвать длиной (*Шесть метров — вполне достаточная длина для коридора*).

Однако таково конституирующее свойство синтаксических дериватов типа S_i. Именно они обладают способностью служить обобщенным названием для одного из актантов исходного слова. Слово *A* является дериватом типа S_i от слова *B*, если *A* может служить обозначением любого *i*-го актанта слова *B*. Как мы уже видели, далеко не для всех слов из упомянутых выше групп такое исходное слово *B*, от которого они могли бы быть синтаксическим дериватом, реально существует. Для слова *вес* существует глагол *весить,* но для многих других параметров, таких, как *длина, ширина, возраст, температура* такого глагола нет. Отсюда следует, что для существа явления S_i не является обязательным, чтобы слова *A* и *B* были разными. Другими словами, можно утверждать, что существительные типа *учитель*, *однополчанин*, *цель*, *длина* являются **актантными дериватами типа S_i от себя самих**, или **актантными автодериватами**. Это не противоречит тому, что эти слова могут оказаться актантными дериватами не только от себя самих, но и от других слов, например, *учитель* от *учить, командир* от *командовать, причина* от *вызывать,* а *вес* от *весить.*

Ниже мы рассмотрим валентные свойства таких слов. В разделе 3 мы обсудим свойства самих автодериватов, а в разделе 4 — выражений адвербиального и адъективного характера, образованных от них (типа *по причине, с целью, под предлогом, длиной, в длину* и т. п.).

3. Актанты автодериватов

Как отмечалось выше, каноническим способом присоединения *i*-го актанта к автодериват типа S_i является присоединение через глагол-связку. В пределах данной статьи мы будем для краткости называть такой актант просто ***i*-м актантом**. Ниже мы рассмотрим некоторые особенности связочной конструкции, образованной автодериватом и его *i*-м актантом, а затем остановимся на активном заполнении валентностей автодериватов.

3.1. Автодериваты и связки

3.1.1. Синтаксическая и коммуникативная структура связочных предложений

С точки зрения синтаксической и коммуникативной структуры связочного предложения можно выделить три класса автодериватов. **Первый класс** образуют абстрактные предикаты типа *причина, цель, подоплека, задача, предназначение* и т. п. Ко **второму классу** относятся параметрические слова типа *высота, длина, ширина, вес, площадь, грузоподъемность* и т. п. Наконец, в **третий класс** входят имена отношений между людьми типа *отец, брат, жених* и т. п.

Начнем с коммуникативной организации связочных предложений. Как для слов класса *причина*, так и для слов класса *высота* в коммуникативно нейтральном контексте типично находиться в теме предложения. Типичная коммуникативная задача для связочных предложений с этими словами — дать ответ на вопрос, какова причина (цель, подоплека и т. п.) некоторого явления или какова длина (ширина, вес и т. п.) некоторого объекта. Так, вторые предложения в следующих парах значительно менее характерны, чем первые:

(16а) *Цельюэтогодемарша* [Th] *быладемонстрация силы* [Rh].

(16б) *?Демонстрация силы* [Th] *была целью этого демарша* [Rh].

(17а) *Высотамачты* [Th] *быладесятьметров* [Rh].

(17б) *?Десять метров* [Th] *была высота мачты* [Rh].

Предложения типа *Десять метров* [Th] *была высота мачты, а не всего корабля,* разумеется, возможны, но коммуникативно маркированы.

Для слов третьего класса, обозначающих отношения между людьми, заметного тяготения ни к тематической, ни к рематической позиции, не наблюдается:

(18а) *Отец Бориса* [Th] *был Леонид* [Rh].

(18б) *Леонид* [Th] *был отцом Бориса* [Rh].

С точки зрения синтаксической позиции автодериват в связочной конструкции распределение несколько другое. В нейтральном контексте для слов типа *причина* и типа *отец* характерно занимать позицию присвязочного члена, в то время, как для слов типа *высота* более нормальна позиция подлежащего:

(19а) *Причинойпожара* [присвяз.член] *былподжог* [подлежащее].

(19б) **Причинапожара* [подлежащее] *былаподжогом* [присвяз. член].

(20а) *ОтцомБориса* [присвяз.член] *былЛеонид* [подлежащее].

(20б) **Отец Бориса* [подлежащее] *был Леонидом* [присвяз. член].

(21а) *Высота мачты* [подлежащее] *(была) десять метров* [присвяз. член].

(21б) **Высотой мачты* [присвяз. член] *были десять метров* [подлежащее].

Здесь следует сделать два замечания.

1. Присвязочный член в предложениях с параметрическими словами типа *высота* не может стоять в творительном падеже: **Высота мачты была десятью метрами*, что затрудняет идентификацию подлежащего в предложениях типа (10а). Тем не менее, исходя из соображений коммуникативной организации (подлежащее тяготеет к тематической позиции) и согласования (связка тяготеет к согласованию с подлежащим) разумно считать, что подлежащим в предложениях типа (10а) является параметрическое существительное, а не обозначение значения параметра.
2. Даже в предложениях со словами классов *причина* и *отец* присвязочный член не всегда стоит в творительном падеже. Так, вполне возможны предложения не только типа *Целью было выжить любой ценой*, *Отцом Бориса был Леонид* но и типа *Цель была выжить любой ценой, Отец Бориса был Леонид*. Здесь решающим соображением оказывается соотносимость именительного падежа с творительным: *цель — целью, отец — отцом*.

3.1.2. Лексический состав глагольного ядра

До сих пор мы говорили о том, что *i*-й актант автодериватов присоединяется к ним через посредство связки *быть*, выступающей либо в полной форме, либо в виде нуля. Однако в этой позиции выступают и другие глаголы, причем некоторые из этих глаголов возможны с одними автодериватами и невозможны с другими. Тем самым связка *быть* предстает как коррелят лексической функции-параметра семейства $Oper_i$ — $Func_i$ — $Labor_{ij}$. Вспомним, что слова типа *причина* выполняют при связке функцию присвязочного члена, в то время как слова типа *высота* — функцию подлежащего. Это значит, что глагол *быть* является значением функции $Oper_2$ от *причина*, и $Func_2$ от *высота*.

Перечислим другие корреляты лексических функций, способные присоединять к автодериватам *i*-й актант.

Прежде всего, допустимы и другие связочные глаголы, такие, как *оказываться* или *становиться*:

(22) *Причина задержки оказалась* ($Func_2$) *в том, что потерялся ключ от зала.*

(23) *С тех пор целью его жизни стало* ($Oper_2$) *разгадать эту загадку.*

С количественными параметрическими словами невозможно употребление глагола *являться*. Ср.:

(24) *Это является* ($Oper_2$) *важной причиной нынешнего кризиса.*

(25) **Высота мачты является десятью метрами.*

Помимо связочных глаголов, в той же лексико-функциональной роли при разных существительных может выступать и ряд других глаголов:

(26) *Это послужило* ($Oper_2$) *предлогом <поводом> к войне <условием для заключения мира, целью данной статьи>.*

(27) *Он доводится <приходится>* ($Oper_1$) *мне братом <дядей, двоюродным дедушкой>.*

(28) *Длина <ширина, вес, грузоподъемность> нового автомобиля равняется половине <составляет половину>* ($Func_2$) *от длины <ширины, веса, грузоподъемности> серийно выпускающейся модели.*

(29) *Координаты Калмыкии составляют* ($Func_2$) (но не: **равняются*) *44°50' и 40°10' восточной долготы и 41°40' и 47°35' северной широты.*

(30) *Причина <цель, результат, условие> состоит <заключается>* ($Func_2$) *в следующем.*

(31) *Французская революция имела* ($Oper_1$) *своей целью* (=ТВОР) *не только изменение прежнего правления, но и уничтожение старой формы общества.*

(32) *Торпеда имела* ($Oper_1$) *длину* (=ВИН) *около 25 метров и вес 40 тонн.*

(33) *Они имеют* ($Labor_{12}$) *в длину всего несколько сантиметров.*

3.2. Активное заполнение валентностей автодериватов, по умолчанию выражающихся разрывно

Мы уже убедились, что дежурным способом реализации той валентности автодериватов, относительно которой уместно рассматривать отношение деривации, является разрывное ее заполнение через связочный (или другой лексико-функциональный) глагол. Обратимся теперь к ситуациям, когда эта валентность нарушает этот канон и все-таки заполняется активным образом. Такая способность обнаруживается у многих предикатных существительных-автодериватов, хотя проявляется она у разных слов по-разному.

Отметим для начала, что практически все интересующие нас существительные имеют возможность присоединить к себе *i*-й актант, образуя с ним аппозитивную конструкцию, в которой последний будет выступать как приложение к нашему существительному, т. е. синтаксически зависеть от него, ср.

(34) *Его отец* (*L*), *генерал (A)*, *был против этого брака*;

(35) *Расстояние* (*L*) *от Москвы до Петербурга (650 км (A)) поезд «Сапсан» преодолевает менее чем за четыре часа*;

(36) *Причину* (*L*) *пожара (поджог* (*A*)*) установили сразу*;

(37) *Этой цели* (*L*) *— обеспечить* (*A*) *бесперебойную работу системы — нам так и не удалось добиться.*

и т. д. Однако активное заполнение валентности предикатного слова с помощью приложения не составляет полноценной альтернативы разрывному заполнению этой валентности через связку: как неоднократно отмечалось (см., например, [3]), конструкции с приложениями вторичны по отношению к связочным конструкциям и семантически очень близки к ним.

Что же касается активного заполнения валентности рассматриваемого вида неаппозитивным способом, то здесь картина оказывается весьма пестрой: возможность/невозможность такого заполнения зависит от лексикографического типа предикатного существительного.

Рассмотрим несколько подробнее один из таких типов — параметрические существительные. Как известно, они имеют единообразную валентную структуру: первый из их аргументов есть предмет, действие или процесс, характеризуемый данным параметром (*длина доски, цвет забора, характер отца*), а другой аргумент (обычно второй) — качественное или количественное значение этого параметра (*длиной в три метра, красный цвет, вспыльчивый характер*).

Параметрические существительные относительно легко допускают активную реализацию валентности значения параметра. В самом частом случае эта валентность оформляется в виде предложно-падежной группы *в* + вин. пад., ср. *Буквально несколько лет назад высота* (*L*) *в* (*A*) *восемь метров казалась непреодолимой*; *Вероятность* (*L*) *в* (*A*) *доли процента для человека — это очень большая вероятность, если речь идет о потере им жизни*; вторичным образом — в виде предложно-падежной группы со значением приблизительности, вводимой предлогами *до, от, около, под, порядка* и некоторыми другими; ср.

(38) *Большой канал, проходящий через весь город [Венецию] и имеющий в длину около 4 км при ширине* (*L*) *до* (*A*) *70 метров*;

(39) *При зарплате* (*L*) *от* (*A*) *50 тысяч до 600 тысяч рублей в год ставка налога составит 13 процентов.*

Допускают активную реализацию соответствующей валентности и автодериваты класса S'_i типа *требование, заявление, жалоба, обоснование* и пр., ср.

(40) *Требованием митингующих была отставка президента — Он выступил с требованием отставки президента.*

По существу, это обстоятельство можно считать конституирующим свойством таких существительных [1].

Термины, обозначающие отношения (*отец, коллега, одноклубник, антагонист* и т. д.), напротив, категорически не допускают активной реализации такой валентности. Так, сочетание типа *#отец Ивана* не может употребляться в ситуации, когда нужно сообщить, что сам Иван является чьим-то отцом.

У существительных-автодериватов других классов способность синтаксически подчинять себе *i*-й актант носит индивидуальный характер. Интересно при этом то, что слова близких семантических классов или даже слова одного семантического класса могут вести себя в этом отношении по-разному. Например, некоторые слова семантического класса ‘причина’ (*причина*, *основание*, *предлог, повод*) практически не допускают активной реализации субъектной валентности или валентности содержания, т. е. той валентности слова *L*, которая может быть названа самим этим словом (о ситуациях, когда такое становится возможным, речь впереди, см. раздел 4). Если, например, слово *причина* используется для характеристики ситуации типа ‘поджог вызвал пожар’, как в предложении

(41) *Причиной пожара явился умышленный поджог здания*,

то словосочетание *#причина поджога* применить нельзя, даже в контекстах, казалось бы, максимально способствующих этому:

(42) *#При расследовании обстоятельств пожара причина поджога рассматривалась в качестве основной.*

Аналогичным образом, если описывается ситуация, когда некто отказался от участия в мероприятии, выдвинув в качестве предлога крайнюю занятость, невозможно употребить выражение *#предлог*

крайней занятости. В то же время такие слова того же семантического класса, как *мотив* или *аргумент*, присоединяют этот актант с большой легкостью, ср.

(43) *В ходе расследования проверялись все возможные мотивы убийцы, в том числе мотив* (*L*) *личной неприязни* (*A*);

(44) *Иногда в поддержку этого тезиса выдвигается спорный аргумент* (*L*) *о* (*A*) *политической пассивности молодежи* или даже … *спорный аргумент* (*L*) *политической пассивности* (*A*) *молодежи.*

Далее, слова *причина* и *цель* принадлежат к одному семантическому ряду; тем не менее, в отличие от *причины,* валентность содержания слова *цель* охотно выражается при нем с помощью синтаксического зависимого:

(45) *Абсурдная цель* (*L*) *создания* (*A*) *вечного двигателя, как это ни прискорбно, находит своих сторонников;*

(46) *На данном этапе Генрих V не ставил цели* (*L*) *создать* (*A*) *объединенное англо-французское королевство.*

Важным классом автодериватов являются существительные, обладающие синтаксическими признаками серии ПРЕД: ПРЕДИНФ, ПРЕДЧТО, ПРЕДЧТОБЫ, ПРЕДВОПР и т. д. (см. [2], [5]). В самом деле, такие признаки приписываются словам, которые способны управлять соответствующими типами языковых единиц через связку (или близкие к ним функциональные глаголы). Например, признак ПРЕДИНФ характеризует слова, которые через связку управляют инфинитивом, причем сам этот инфинитив является подлежащим; ср.

(47) *Самой большой проблемой* (=прединф) *было уговорить директора*.

Признаки ПРЕДЧТО, ПРЕДЧТОБЫ и ПРЕДВОПР характеризуют слова, которые через связку управляют придаточным-подлежащим, вводимым, соответственно, союзами *что* или *чтобы*:

(48) *Для меня было новостью* (=предчто)*, что Трир такой древний город;*

(49) *Единственное направление движения не является достаточным основанием* (=предчтобы)*, чтобы не пользоваться указателями поворота*).

Наконец, признак ПРЕДВОПР характеризует такие слова, которые через связку управляют придаточным-подлежащим в виде косвенного вопроса; ср.

(50) *Куда идет король — большой секрет* (=предвопр);

(51) *Остается загадкой* (=предвопр)*, есть ли жизнь на других планетах.*

Очевидно, что все слова или выражения, выступающие в предложениях (36)–(40) в роли подлежащего, являются *i*-ми актантами существительных, имеющих указанные синтаксические признаки. Способность таких неименных актантов синтаксически подчиняться автодериватам тоже носит в целом индивидуальный характер. Так, по нашим оценкам, из 170 существительных, которым в русском комбинаторном словаре лингвистического процессора ЭТАП-3 [3] приписан признак ПРЕДИНФ, лишь около 20 допускают активное заполнение соответствующей валентности инфинитивом: это такие слова, как *желание* (ср. *Моим желанием было немедленно уйти — У меня появилось желание немедленно уйти*), *обязанность (учитывать интересы всех), мысль (поехать на Кавказ), идея (пригласить гостей), необходимость (зарабатывать на хлеб), перспектива (остаться без денег)* и др. под. В то же время для подавляющего большинства таких слов активное присоединение *i-го* актанта невозможно, ср. **подвиг вступиться за коллегу, *наглость просить надбавку в кризис, *преступление уклоняться от уплаты налогов* и т. д.

Следует особо отметить, что рассматриваемые валентности автодериватов могут активно заполняться не только посредством именных или предложно-падежных групп, как во всех приведенных выше примерах, но и посредством согласующихся с ними прилагательных, ср. например, *Длина* (*L*) *волны — 3 сантиметра* (*P*) и *трехсантиметровая* (*P*) *длина* (*L*) *волны.*

В некоторых случаях адъективную реализацию интересующей нас валентности допускают даже автодериваты, сопротивляющиеся активному заполнению валентностей именами и предложными группами. Мы уже детально рассматривали валентные свойства слова *причина* и видели, что оно отвергает активную реализацию субъектной валентности. Этот запрет не распространяется, однако, на прилагательные, во всяком случае, местоименные. Посмотрим, например, на конструкции, когда местоимением заполняется объектная валентность слова *причина*:

(52) *Причиной пожара явился поджог — Причиной* ***этого*** *явился поджог — Причиной* ***чего*** *явился поджог? — Произошел пожар, причиной* ***которого*** *<причиной* ***чего****> явился поджог*

и т. д.

В противовес этим выражениям, субъектную валентность слова *причина* способны заполнять согласующиеся местоименные прилагательные:

(53) *Какую* (*P*) *причину (L) рассматривали в качестве основной при расследовании обстоятельств пожара?,*

(54) *Эта* (*P*) *причина (L) рассматривалась в качестве основной* и т. д.

В этой связи обращают на себя внимание контрастные пары типа *причина этого — эта причина,* требующие противоположной интерпретации, каковая до сих пор, насколько известно авторам, специально не обсуждалась.

Разумеется, все такие особенности заполнения валентностей должны найти отражение в толково-комбинаторном словаре.

4. Адвербиальные и адъективные производные автодериватов.

Рассмотрим теперь соотношение автодериватов с выражениями, являющимися по существу их синтаксическими производными. Мы имеем в виду такие предложно-падежные и падежные формы, которые по смыслу близки к «оглаголенному» предикатному существительному-автодериват в адвербиальной или адъективной форме, — деепричастному или причастному обороту от соответствующих предикатов (т. е. от значений лексических функций типа Oper-Func-Labor для наших существительных — *быть, иметь* и др.)

Примеры таких производных можно увидеть в предложениях типа

(55) *Котел взорвался по причине короткого замыкания* (≈ ‘котел взорвался, имея причиной короткое замыкание’);

(56) *Акция проведена с целью устрашения оппозиции* (≈ ‘устрашение оппозиции будучи целью проведения акции’);

(57) *Перемирие будет заключено при условии отказа от насилия* (≈ ‘отказ от насилия будучи условием заключения перемирия’);

(58) *Самолет летел со скоростью 900 км/час* (‘самолет летел, имея скорость *900 км/час*’);

(59а) *Здесь требуются доски длиной 3 м* (‘… доски, имеющие длину <в длину> 3 м’);

(59б) *Здесь требуются доски трехметровой длины*;

(60) *В гарнизоне можно было разместить роту численностью до 200 человек с полным вооружением.* (‘… роту, имеющую численность до 200 человек’)

Как мы видим, предложно-падежные формы этих производных отличаются достаточно большим разнообразием: могут использоваться различные предлоги (*по причине*, *с целью* и *в целях*, *при условии* и *на условиях*, *по привычке, на высоте, со скоростью, по имени* и *под именем*) и т. д., а также, по крайней мере, родительный и творительный падежи существительного (*длиной в три метра* и *трехметровой длины*).

Все эти выражения по существу представляют собой синтаксические фраземы, описание которых остается за рамками нашего изложения.

Мы отметим здесь лишь одно явление, представляющееся достаточно типичным.

Если сравнить употребление слова типа *причина* в независимом окружении и в составе адвербиала типа *по причине,* может показаться, что они имеют разные валентные структуры — у слова *причина* реализуется в первую очередь объектная валентность (*причина пожара*), а у адвербиала *по причине* имеется одна субъектная валентность *(по причине поджога)*. По нашему мнению, это не так. И у *причины,* и *у по причине* набор валентностей один и тот же — у обеих единиц есть и субъектная, и объектная валентности. Различие здесь лишь в способе реализации этих валентностей. Как мы видели, у независимой *причины* объектная валентность реализуется активно, а субъектная — разрывно, через ЛФ-глагол. Что касается адвербиала, то здесь имеет место обратная ситуация: активно реализуется субъектная валентность слова *причина* **в составе** синтаксической фраземы *по причине,* а ее объектная валентность реализуется разрывно. Так, в предложении (44) эта валентность выражена подлежащим главного глагола *котел*, не имеющим непосредственной синтаксической связи с адвербиалом *по причине.*

5. Заключение

Было показано, что обширный класс слов, относящихся к разным тематическим группам, обладает тем не менее интересным общим свойством: все эти слова могут служить обозначением одного из своих актантов. Эта способность реализуется прежде всего в связочных предложениях, хотя и не только в них: *Иван — мой однополчанин. Длина доски — три метра. Причина взрыва — короткое замыкание. Эта длина* (т. е. три метра). *Такая причина* (т. е. короткое замыкание).

Один из подклассов этого класса — отглагольные актантные дериваты типа S_i — достаточно хорошо изучен. Например, *учитель* и *ученик* являются субъектным и объектным дериватом глагола *учить*.

За этими существительными стоит ситуация, в которой есть и обучающий, и обучаемый. Если Иван учит Петра, то и Иван, и Петр заполняют валентности как глагола *учить,* так и существительных *учитель* и *ученик*. Оказывается, что способность служить обозначением собственного актанта свойственна далеко не только дериватам от глаголов, но и многим другим словам, которые мы предложили называть автодериватами.

Ключевым фактором являются не связочные предложения сами по себе — в них легко участвуют существительные любых классов; ср., например, *Береза — дерево*. Существенное обстоятельство состоит в том, что связка оказывается средством присоединения актанта.

Были рассмотрены синтаксические и коммуникативные свойства связочных предложений, присоединяющих актант к автодериватам. Хотя наиболее типичным глагольным ядром подобных предложений является глагол *быть*, тем не менее здесь наблюдается определенное лексическое и синтаксическое разнообразие, заставляющее признать, что глагол *быть* является лишь одним из коррелятов семейства лексических функций Oper_i — Func_i — Labor_{ij}.

Особого внимания заслуживают адвербиальные и адъективные дериваты автодериватов — выражения типа *по причине, с целью, на высоте, под предлогом, длиной* и т. п. Некоторые их актантные свойства прямо противоположны актантным свойствам исходных слов.

Литература

1. *Апресян Ю. Д.* Лексическая семантика. Синонимические средства языка. М., Наука, 1974.
2. *Апресян Ю. Д., Богуславский И. М., Иомдин Л. Л., Лазурский А. В., Митюшин Л. Г., Перцов Н. В., Санников В. З., Цинман Л. Л.* Лингвистический процессор для сложных информационных систем. М.: Наука, 1992.
3. *Богуславский И. М.* Сфера действия лексических единиц. М.: Школа «Языки русской культуры», 1996.
4. *Igor Boguslavsky.* On the Passive and Discontinuous Valency Slots, Proceedings of the 1st International Conference on Meaning-Text Theory. Paris, Ecole Normale Supérieure, June 16–18; 2003.
5. *Иомдин Л. Л., Мельчук И. А., Перцов Н. В.* Фрагмент модели русского поверхностного синтаксиса. I. Предикативные синтагмы // Научно-техническая информация. Серия. 2. № 7. 1975, С. 30–43.

Ассоциативная сеть понятий, образующих запросы к интернету

An associative network of concepts that enter to internet queries

Большаков И. А. (bolshakov34@mail.ru),
Независимый исследователь, Москва

Большакова Е. И. (bolsh@cs.msu.ru),
Московский Государственный Университет, Москва

Гельбух А. Ф. (gelbukh@gelbukh.com),
Национальный Политехнический Институт, Мехико, Мексика

В базе пользовательских запросов поисковиков Google и Яндекс выявлена обширная совокупность сочиненных пар существительных и на их основе построена и описана ассоциативная сеть понятий, из которых часто формируются русскоязычные запросы к Интернету. Показано, что выявленные пары существительных представительно входят и в текстовые массивы Интернета. Исследована полученная ассоциативная сеть.

1. Введение

Уже более 10 лет интернетовские поисковые машины типа Google накапливают базу пользовательских запросов (БПЗ) для своей внутренней надобности — совершенствования поискового аппарата. Запрос может состоять из одного или нескольких слов, быть грамматически связанным (*производство дверей*) или несвязанным (*кирпич цена*), орфографически безупречным или содержащим ошибку.

Пару лет назад Google позволил пользователям оперативно общаться с этой БПЗ. В процессе посимвольного ввода пользователем запроса поисковик выводит и модифицирует меню из упорядоченных подсказок, начальные символы которых совпадают с уже введенной частью запроса. Часто можно не вводить свой запрос целиком, а лишь выбрать одну из строк образующегося меню.

Обнаружилось, что при запросе с более чем одним понятием пользователи, не прибегая к средствам расширенного поиска с формальными логическими связками, часто соединяют понятия союзом *и*. Например, если беременная пользовательница хочет оценить для себя опасность аллергии, она вводит запрос *беременность_и_*, (_ означает пробел), а затем выбирает в образовавшемся меню строку *беременность и аллергия*. При этом Google сообщает, что таких запросов уже было 515 тыс., а это значит, что запрос привычен для поисковика и на него есть релевантные ответы. Доступ к БПЗ сейчас есть и у русских поисковиков (Яндекс, Рамблер). При этом русская БПЗ Яндекса оказалась побогаче, чем у Google и Рамблера.

Тем самым современные поисковики предоставляют прикладному лингвисту среди многого прочего тысячи сочиненных пар указанного вида. Возникает возможность создать или резко пополнить коллекцию сочиненных именных пар. Это не только интересно с языковой точки зрения, но и позволяет создать ассоциативную сеть понятий, которыми оперирует русскоязычный пользователь Интернета. В части ассоциаций здесь никаких методологических проблем не возникает, но в принадлежности некоторых сочиненных пар из БПЗ русскому языку в целом можно и усомниться. Действительно, можно ли считать принадлежащими русскому языку такие пары, как *ноги и характер* (2,49 млн. запросов к Google, январь 2010 г.) или *ангина и керосин* (12,1 тыс. запросов)?

Данная работа преследует следующие цели:

- Описать коллекцию сочиненных именных пар, входящую в словарь КроссЛексика [1], после существенного пополнения ее данными из Google и Яндекса;
- Дать приближенную интерпретацию ряда характерных запросов из сочиненной пары, по-

казав на примерах несводимость возникающих ассоциаций к семантическим связям типа WordNet [3];
- На основе статистических данных показать, что новые пары понятий встречаются и на сайтах Интернета, а потому могут считаться принадлежащими русскому языку в целом;
- Построить из компонентов сочиненных пар ассоциативную сеть понятий, которыми оперирует русскоязычный пользователь в запросах к Интернету;
- Проанализировать построенную сеть, выявив понятия с максимальным количеством ассоциативных связей, вычленив и описав связные компоненты сети и др.

2. Пополненная коллекция сочиненных пар

В коллекции сочиненных пар словаря КроссЛексика [1, 2] до привлечения содержимого БПЗ было 4,1 тыс. пар, бытующих в текстах и подтвержденных сайтами Интернета. Из них субстантивных пар было 2,8 тыс., т. е. 68%. Примерно в 10 процентах случаев компоненты субстантивных пар были не одиночными существительными, а группами типа *авторские права, автоматическое оружие* или *железные дороги*.

Семантически компоненты пар соотносились по одному из следующих вариантов:
- когипонимы в некой родовидовой иерархии (*руки и ноги, аксиомы и теоремы, труд и капитал, акушерство и гинекология*);
- синонимы, квазисинонимы и повторы (*траур и скорбь, горести и несчастья, тысячи и тысячи*);
- антонимы, квазиантонимы, противоположные понятия и конверсивы (*бедные и богатые, актив и пассив*, *Бог и дьявол*, *купля и продажа*, *действие и противодействие*);
- парные названия и исторически связанные имена (*Босния и Герцеговина, Адам и Ева*).

Лишь изредка компоненты оказывались соучастниками некоторой ситуации (*писатель и читатели, закон и порядок, кожа и косметика*) или понятиями, связанными причинно-следственными связями (*война и разруха, преступление и наказание*, *штормы и наводнения*).

Для целей данной работы мы брали пары не только из Google, но и из Яндекса, поскольку статистического порога извлечения новых пар из БПЗ не устанавливалось. В первом туре пополнения в качестве первого компонента пары брались субстантивы *S*, уже содержащиеся в коллекции в качестве первого либо второго компонента. Вручную вводились запросы вида *S_и_*. В окне поисковика возникали различные пары, частично новые для коллекции. Последними и велось пополнение. При этом в качестве второго компонента в новых парах часто появлялись и новые понятия. Они брались началом аналогичных запросов во втором туре, что выявляло новые компоненты на второй позиции в паре. В третьем туре новые вторые компоненты брались первыми компонентами очередных запросов и т. д. После семи туров пополнения общее количество пар достигло 16 тыс., а число субстантивных пар превысило 14 тыс. (более 90 % всей коллекции), увеличившись в 5,6 раз.

Чтобы обеспечить смысловую самостоятельность обоих компонентов пар, мы не брали довольно частые пары с анафорической отсылкой *его/ее/их* во втором компоненте (например, *кризис и его последствия*), а также пары типа *газеты и другие/иные СМИ*. Были отброшены повторы. Замеченные редкие синонимы, иногда — абсолютные (*Белоруссия = Беларусь, кормление грудью = грудное вскармливание, зарплата = заработная плата*) не объединялись.

Смысл сочинения компонентов-понятий в парах стал после пополнений много разнообразнее. Охарактеризуем это на примерах. Так, множество пар *X и цены* эквивалентно по смыслу *цены*(*X*)?, т. е. пользователь просто заменил в запросе подчинительную связь сочинительной. Это не касается пар *цены и комплектация/наличие/скидки/ценообразование*, где пользователь интересуется разными параметрами некоторого объекта, непосредственно парой не отражаемого (см. табл. 1).

Более сложен смысл сочинения у образцов запросов к Google, представленных в табл. 1. Например, для большинства пар, включающих *Y = беременность / здоровье*, семантическим представлением запроса является *влияние*(X,Y)? первого аргумента на второй, например, *влияние*(*простуда*, *беременность*)?. Для пар, включающих *СМИ*, запрос примерно эквивалентен симметричному предикату *взаимодействие*(X,Y)?, например, *взаимодействие*(*СМИ*, *власть*)? В парах, где X = *йога,* а Y = *православие / христианство / буддизм*, замечаем симметричный предикат *совместимость*(*йога*,Y)? Осмысление пары *ангина и керосин* позволяет предположить структуру уже с двумя предикатами: *эффективность*(*лечение*(*ангина*, *керосин*))?. Подобные толкования разного уровня сложности могут быть предложены и для всех иных запросов.

Скрытая семантика запросов, как видим, разнообразна, а компоненты пар связаны неизвестным предикатом, и быть может, не одним. Эти предикаты может предположить лингвист, но алгоритмически они в общем случае не восстановимы. Поэтому поиск простых и коротких семантических связей между компонентами пары {X, Y} средствами, например, WordNet [3], который оперирует синонимами, антонимами, гипонимами / гиперонимами и меронимами / холонимами, в общем случае без-

Таблица 1. Статистика образцов запросов и ответов
(VQ — число запросов, VS — число прямых ответов, VF — число косвенных ответов, все в тысячах)

п/п	Сочиненная пара	VQ	VS	VF
1	беременность и роды	1470,0	1380,0	1720,0
2	беременность и простуда	219,0	249,0	263,0
3	беременность и компьютер	784,0	99,2	834,0
4	беременность и месячные	271,0	201,0	251,0
5	беременность и курение	494,0	52,0	499,0
6	беременность и питание	1450,0	37,8	1470,0
7	беременность и грипп	460,0	258,0	593,0
8	беременность и молочница	171,0	125,0	163,0
9	здоровье и красота	99700,0	2110,0	144000,0
10	здоровье и материнство	108,0	118,0	195,0
11	здоровье и спорт	315000,0	173,0	261000,0
12	здоровье и комфорт	915,0	178,0	926,0
13	здоровье и здоровый образ жизни	1960,0	81,6	1180,0
14	здоровье и долголетие	243000,0	40,1	310000,0
15	здоровье и окружающая среда	558,0	121,0	426,0
16	здоровье и питание	266000,0	455,0	495000,0
17	здоровье и компьютер	158000,0	21,2	269000,0
18	СМИ и периодика	249,0	124,0	202,0
19	СМИ и политика	9670,0	54,9	10400,0
20	СМИ и власть	3690,0	5,8	3570,0
21	СМИ и культура речи	561,0	13,3	315,0
22	СМИ и общество	8930,0	561,0	10300,0
23	СМИ и общественное мнение	2180,0	39,4	1350,0
24	СМИ и PR	2860,0	364,0	2840,0
25	СМИ и государство	14100,0	298,0	8520,0
26	СМИ и выборы	2300,0	116,0	986,0
27	СМИ и культура	8550,0	203,0	9700,0
28	ангина и беременность	64,8	34,0	67,1
29	ангина и антибиотики	181,0	1,2	139,0
30	ангина и кашель	584,0	42,1	1110,0
31	ангина и гомеопатия	50,6	0,0	18,9
32	ангина и грудное вскармливание	240,0	2,2	767,0
33	ангина и керосин	12,0	0,0	5,8
34	ангина и сердце	5010,0	6,3	7300,0
35	ноги и характер	2490,0	198	2370,0
36	алгоритмы и структуры данных	359,0	10,2	252,0
37	алгоритмы и исполнители	80,6	10,8	69,2
38	алгоритмы и программы	720,0	160,0	592,0
39	алгоритмы и процессоры	63,1	46,2	170,0
40	алгоритмы и блок-схемы	96,8	1,5	94,0
41	аккумуляторы и зарядные устройства	3230,0	97,1	2320,0
42	аккумуляторы и цены	2500,0	0,0	2290,0
43	йога и питание	20800,0	1200,0	32300,0
44	йога и православие	197,0	23,1	138,0
45	йога и беременность	290,0	97,1	372,0
46	йога и здоровье	45200,0	244,0	48200,0
47	йога и похудение	253,0	1,2	143,0
48	йога и алкоголь	157,0	11,1	151,0
49	йога и христианство	186,0	9,3	194,0
50	йога и сколиоз	10,1	1,1	16,3

надежен. Но для десятков, сотен тысяч или даже миллионов пользователей Интернета эти компоненты очевидно связаны по смыслу, а потому введение в оборот компьютерной лингвистики отдельной семантической подсистемы таких связей представляется оправданным.

Специфичность сочинения в запросах позволяет усомниться в наличии некоторых подобных пар в обычных текстах языка. Для прояснения ситуации мы произвольно взяли из Google 50 пар со статистикой БПЗ (столбец *VQ* табл. 1), а также измерили число страниц в основной базе данных Google, которые включают те же пары буквально (т. е. контактно и в номинативе, столбец *VS*) или на произвольном расстоянии внутри абзаца и с произвольным падежом (столбец *VF*). Все числа измерялись в тысячах. Будем называть столбец *VS* статистикой прямых ответов, а столбец *VF* — косвенных ответов. За исключением трех случаев (6%), сайты Google содержат прямые ответы, и среднее отношение числа прямых ответов к числу запросов составляет $k_{SQ} = 0{,}19$. Косвенные ответы встречаются на сайтах много чаще, со средним коэффициентом пропорциональности $k_{FQ} = 1{,}1$.

Поскольку конкретное значение коэффициента k_{SQ} резко меняется от пары к паре, вычислим еще косинусы углов между векторами *VQ, VS, VF* с координатами $i = 1\ldots50$. Косинус угла между *VQ* и *VS* дается формулой

$$COS(VQ,VS) = \frac{\sum_i \left(VQ_i\, VS_i\right)}{\sqrt{\sum_i \left(VQ_i\right)^2 \sum_i \left(VS_i\right)^2}},$$

а для прочих пар — аналогичными формулами, что дало $COS(VQ,VS) = \mathbf{0{,}26}$; $COS(VQ,VF) = \mathbf{0{,}95}$; $COS(VS,VF) = \mathbf{0{,}27}$. Иными словами, вектора запросов и косвенных ответов практически коллинеарны, но различие между векторами запросов и прямых ответов существенно. Поэтому запросы в виде пар могут считаться принадлежащими русскому языку лишь в своем большинстве.

3. Семантическая сеть запросов

После пополнения коллекции сочиненных пар был программно построен неориентированный граф с вершинами, помеченными различными понятиями. Ребра графа соединяют вершины с понятиями X, Y, если последние входят в пару *X и Y* и/или *X и Y*. Например, вершина *ангина* связана с вершинами *антибиотики, беременность, гомеопатия, грудное вскармливание, кашель, керосин, мороженое, прополис, сердце2, фарингит*. Здесь и далее цифра в конце слова означает номер омонима.

Таблица 2. Степени *D* наиболее популярных понятий

D	**Понятие**	*D*	**Понятие**
302	*беременность*	31	*ремонт*
110	*здоровье*	29	*армия*
87	*алкоголь*	29	*методы*
87	*цены*	29	*экономика*
54	*спорт*	28	*давление1*
52	*культура1*	28	*лечение*
51	*похудение*	28	*функции1*
49	*дети*	27	*безопасность*
48	*человек*	27	*власть*
41	*диабет*	27	*реклама*
40	*диета*	27	*экология*
39	*курение*	26	*структура1*
39	*любовь*	25	*философия*
37	*общество1*	24	*контроль*
37	*религия*	24	*Наука*
37	*Россия*	24	*пиво*
36	*право2*	24	*христианство*
34	*температура*	23	*Водка*
34	*характер2*	23	*государство*
33	*бизнес*	23	*деньги*
33	*дизайн*	23	*Интернет*
32	*кризис*	23	*искусство*
32	*развитие*	23	*православие*
31	*политика*	23	*прыщи*

Полученный граф назовем ассоциативной сетью [4]. На март 2010 г. она включала 9,2 тыс. вершин-понятий с общим числом ребер-ассоциаций 25,2 тыс. (в среднем 2,73 ребер на вершину). В таблице 2 дана степень *D* (число ребер) наиболее популярных понятий. Два первых места с большим отрывом занимают *беременность* и *здоровье*. Далее устойчиво расположились *алкоголь, цены, спорт, культура1, похудение, дети* и *человек*. Значит, все это наиболее острые проблемы для пользователей. *Алкоголь* поднимается на одну позицию, если присоединить *пиво*, *водка*, *алкогольные напитки* и *алкоголики*. Но все, прямо связанное со здоровьем (*беременность, здоровье, спорт, похудение, диабет, диета* и др.), по совокупности всегда первенствует. Среди лидеров всегда остаются также *цены*, но они интересны далеко не для всех товаров и услуг, вошедших в сеть.

Таблица 3. Степени *D* популярных многословных понятий

D	**Понятие**
22	*окружающая среда*
20	*щитовидная железа*
16	*кормление грудью*

D	**Понятие**
14	*лунный календарь*
13	*грудное вскармливание*
12	*социальная политика*
12	*характерные черты*
11	*государственное управление*
11	*группа крови*
11	*международное право*
11	*охрана окружающей среды*
10	*охрана природы*
10	*Рынок труда*
10	*экономический рост*
9	*витамин С*
9	*глобальные проблемы*
9	*заработная плата*
9	*культура речи*
9	*Новый год*
9	*общественное мнение*
9	*социальный контроль*
8	*бронхиальная астма*
8	*зеленый чай*
8	*знаки зодиака*
8	*информационные технологии*
8	*образ жизни*
8	*оливковое масло*
8	*охрана труда*
8	*рыночная экономика*
8	*социальная справедливость*
8	*тепловые двигатели*
7	*валютный курс*

Многословные понятия представлены в сети достаточно широко. В таблице 3 даны наиболее популярные из них. Интерес пользователей к окружающей среде и ее охране, щитовидной железе, кормлению грудью, группе крови, рынку труда, витамину С и заработной плате кажется естественным. Удивляет, однако, высокая популярность лунного календаря, выдающая интерес широкой массы пользователей к эзотерике. Отметим еще популярность понятий социальной политики, государственного управления, международного права, экономического роста, информационных технологий, тепловых двигателей и др., что, похоже, отражает интерес студентов разных специальностей к материалам для рефератов и курсовых, добываемым из Интернета.

Изучим теперь связность созданной ассоциативной сети: распадается ли она на изолированные компоненты связности (подсети), и если да, то каковы эти подсети? Путем программного обхода вершин были найдены все компоненты связности и рассчитаны их параметры (см. табл. 4). **Мощностью** графа называется число узлов в нем [4]; **диаметром** — самая длинная из кратчайших цепей, связывающих какие-либо две вершины графа; **мостом** — ребро, разрыв которого увеличивает число компонентов связности графа; **точкой сочленения** — вершина, удаление которой ведет к увеличению числа компонентов связности; **степенью вершины** — число ребер, которым она принадлежит; **висячей вершиной** — вершина степени 1.

Всего в сети (на январь 2010 г.) было 870 связных подсетей. Доминирующая подсеть содержала 5129 вершин. Ее тематика предельно широка, мы назвали ее общежитейским универсумом. Сюда вошли все понятия таблиц 2 и 3. С огромным отрывом далее шла подсеть о воспитательно-образовательной сфере с 21 вершиной. На третьем месте — две подсети по 13 вершин, отражающих детали домов и преступность соответственно. На предпоследнем по мощности месте оказалось 117 подсетей из трех вершин, на последнем месте — 654 подсети из двух вершин. Тематика подсетей из 2–5 вершин узка и многообразна, мы обозначили ее как <фрагменты>. Можно заметить, что диаметры часто включают мосты, а мосты включают многие точки сочленения.

Особый интерес вызывает доминирующая подсеть. Ее размеры столь велики, что программное (кубической сложности) вычисление ее диаметра оказалось неосуществимым даже за неделю счета. Но анализ показал, что 2836 вершин этой подсети (т. е. более половины всех вершин) являются висячими. Подсеть-доминанта похожа на ежа с громадным количеством торчащих иголок типа *абажуры, аббревиатуры, абитуриенты, абсурд, авансирование, автоаксессуары, автокосметика, автокредит, авторизация, авторитаризм, автохимия*. Почти все эти понятия редки в обиходе. Если отделить висячие вершины, оставшаяся более компактная и более связная подсеть из 2208 вершин уже позволяет вычислить ее диаметр за ограниченное время (снабжен знаком + в таблице 4).

Полезно заметить, что пополнение сети систематически ведет к слиянию малых подсетей с доминантой. Так, появление ассоциации *ложки — столовые приборы* вызовет присоединение к доминанте подсети из восьми слов (см. таблицу 4), поскольку в доминанте уже содержится вершина *столовые приборы*. Эта тенденция не исключает появление небольших изолированных подсетей при будущих пополнениях, и до некоторого этапа они могут разрастаться отдельно, но потом сольются с доминантой. При всем этом всегда сохранится множество двухвершинных подсетей типа *Адам — Ева*. Для входящих в них понятий трудно вообразить иные ассоциации, выразимые сочиненными парами.

Образование всеподавляющей доминантной подсети внутри сформированной полной сети вполне коррелирует с результатами тех исследований по семантике слов средствами WordNet [5], которые основаны на теории случайных графов [6, 7].

Таблица 4. Связные подсети и их характеристики

Мощность	Подсетей	Длина диам.	Примеры диаметров	Примеры мостов	Примеры тчк. сочленения	Тематика
5129	1	14+	продавцы–покупатели–поставщики–закупки–снабжение–комплектация–цены–ламинанат–вода–ветер–снег–грозы–дожди;	гололедица–снег; комплектация–цены; питательные вещества–вода;	цены; снег; вода; водка;	общежитейский универсум
21	1	10	любители–профессионалы–дилетанты–специалисты–ЕГЭ–вузы–школы1–колледжи–лицеи–гимназии;	***специалисты–ЕГЭ***; ***ЕГЭ–вузы; лицеи–гимназии; ясли–детсады***;	вузы; ЕГЭ; детсады; институты; специалисты	воспитательно-образовательная сфера
13	2	7	фасад–кровля–фасады–кровли–крыша1–перекрытия–пустоты;	изоляция–кровли; кровли–фасады;	кровли; крыши; перекрытия;	(1)детали домов (2)преступность
11	3	6–8	диаметр–окружность–круг1–крест–шар–сфера1;	диаметр– окружность; круг1–крест;	крест; круг1; окружность; шар	(1)геометр. фигуры (2)фазы изменения (3)стройматериалы
10	3	5–7	мародерство–грабежи–убийства–теракты– диверсии;	диверсии–теракты; пытки–убийства;	грабежи; теракты; убийства	(1)преступные акты (2)выч.математика (3)лингвистика
9	2	3, 4	БМП–танки1–артиллерия–ПВО;	БМП–танки1; артиллерия–ПВО;	артиллерия танки1	(1)боевая техника (2)студенч.работы
8	7	3–6, 8	***перила–лестницы–лифты–подъемники; перчатки–шарфы– палантины;***	лестницы–лифты; лифты–подъемники; варежки–шарфы;	лестницы; лифты; шарфы	(1)лестницы-лифты (2)одежда (3)оружие (4)насекомые (5)термодинамика (6)деды-внуки (7)авторы-герои
7	4	3, 4	хранение файлов–хостинг–домен–рабочая группа;	хостинг–домен;	хостинг; домен	(1)сайты-домены (2)персон. черты (3)покрытия (4)волны-частоты
6	9	4–6	соблазны–уговоры–обещания–декларации; операционные системы –утилиты–драйверы–загрузка–выгрузка– погрузка;	обещания–уговоры; выгрузка–загрузка; драйверы–загрузка; утилиты–драйверы;	***обещания; уговоры; выгрузка; драйверы; загрузка;***	(1)уговоры (2)компьютеры (3)физика (4)сады-парки (5)лингвистика (6)изменения (7)медучреждения (8)охота (9)оружие
5	22	3–5	беспризорники–сироты–вдовы–вдовцы;	беспризорники–сироты; вдовы–вдовцы;	вдовы; сироты;	<фрагменты>
4	45	3–4	отличие–сходство–различие–повторение;	отличие–сходство; различие– повторение;	различие; сходство;	<фрагменты>
3	117	3	Тангаж–крен–дифферент;	тангаж–крен; крен–дифферент;	крен;	<фрагменты>
2	654	2	пассив–актив1;	пассив–актив1;	не приложимо	<фрагменты>

4. Заключение

Из баз пользовательских запросов поисковиков Google и Яндекс извлечена обширная совокупность сочиненных именных пар и на этой основе построена и описана ассоциативная сеть из 9,2 тыс. понятий, которые обычно входят парами в запросы к Интернету. У каждого из этих понятий в среднем по 2,7 ассоциативных партнеров. Тем самым предложена новая семантическая подсистема, базирующаяся на тех смысловых ассоциациях между понятиями, которые в большинстве не сводимы к связям типа WordNet.

Показано, что выявленные именные пары обычно входят с достаточной частотой и в текстовые массивы Интернета и потому, за редкими исключениями, могут считаться принадлежащими русскому языку в целом.

Созданную ассоциативную сеть удобно использовать при оперативном построении запросов к тем поисковикам, которые либо пока не содержат данные пары понятий в своей базе пользовательских запросов, либо вообще не предоставляют пользователю доступ к этой базе. Можно также использовать эту сеть в программах классификации и установления семантической связности текстов.

Полезно было бы сравнить выявленные ассоциации с теми, которые известны по русскому ассоциативному словарю [8] и по идеографическому словарю типа тезауруса Роже [9]. Нечто интересное может выявиться при применении к полученной сети известных методов кластеризации. Когда станет известной аналогичная ассоциативная сеть для английского языка, можно будет сравнить интересы русскоязычных и англоязычных пользователей. Содержимое предложенной сети выложено на сайте [10].

Настоящая работа выполнена при частичной поддержке второго автора правительством Мексики (гранты SNI, SIP-IPN 20100773 и CONACYT 50206-H.)

Литература

1. *Большаков И. А.* КроссЛексика — большой электронный словарь сочетаний и смысловых связей русских слов. // Компьютерная лингвистика и интеллектуальные технологии. Международная конференция «Диалог 2009». Вып. 8 (15) М.: РГГУ, 2009. С. 45–50.
2. *Bolshakov I. A.* Stable Coordinated Pairs as a Specific Resource of Language. // East West Encounter: Second Intern. Conf. Meaning-Text Theory MTT'2005. Slavic Culture Languages Publ., Moscow, 2005, p. 86–97.
3. *Fellbaum Ch. (Ed.)* WordNet: An Electronic Lexical Database. // MIT Press, 1998.
4. *Leman F.* Semantic Networks. // Computer Math. Applic. Vol. 23, No. 2–5, 1992, p. 1–50.
5. *Kamps J., Marx M., Mokken R. J., de Rijke M.* Using WordNet to Measure Semantic Orientations of Adjectives. // Proceedings of the Fourth International Conference on Language Resources and Evaluation LREC2004, vol. IV, 2004, p. 1115–1118.
6. *Erdos P., Renyi A.* On the evolution of random graphs. // Magyar Tudomanyos Akademia Matematikai Kutato Intezetenk Kozlemenyei. No.5, 1960, p. 17–61.
7. *Janson S., Luczak T., Rucinski A. Random Graphs. // NY: Wiley — Interscience Series in Discrete Mathematics, 2000.*
8. *Караулов Ю. Н., Сорокин Ю. А., Тарасова Е. Ф., Уфимцева Н. В., Черкасова Г. А. Русский ассоциативный словарь. // М.: АСТ, 1994.*
9. *Баранов О. С. Идеографический словарь русского языка. // М.: ЭТС, 1995.*
10. *Ассоциативная сеть русскоязычных пользователей Интернета.* www.Gelbukh.com/coordinate-pairs-in-queries/

Обучение классификаторов на основе выделения фрагментов

Classifier training by passage recognition

Васильев В. Г. (vvg_2000@mail.ru)
ЛАН-ПРОЕКТ

В данной работе рассматривается новый метод обучения классификаторов, основанный на использовании результатов автоматического выделения фрагментов в текстах. Также приводится описание известных и нового метода выделения фрагментов в текстах. Проводятся эксперименты со стандартными тестовыми массивами, как на русском, так и на английском языке.

1. Введение

В настоящее время вопросы построения эффективных средств автоматической классификации текстов достаточно активно рассматриваются как в отечественных, так и в зарубежных работах. При этом в последнее время наибольшее распространение для построения классификаторов получил подход, основанный на использовании статистических или обучаемых методов. Данные методы опираются на использование эталонных массивов документов, в которых документы вручную разложены экспертами по заданным рубрикам, для автоматического построения статистических моделей и правил классификации.

Необходимо отметить, что в большинстве работ в области обучаемой классификации документы рассматриваются как неделимые цельные сущности и представляются в виде вектора весов информационных признаков. При таком подходе внутренняя структура документов и последовательность следования слов не учитывается, что является приемлемым при обработке относительно небольших или тематически однородных документов. В случае, когда документы являются политематическими или имеют сложную внутреннюю структуру, данный подход не подходит. В этой ситуации в ряде работ предлагается использовать подходы, основанные на поиске и классификации фрагментов в текстах. Рассмотрим их более подробно методы поиска фрагментов и методы классификации фрагментов, которые описываются в современной литературе.

Задача поиска фрагментов заключается в извлечении частей документа, которые релевантны запросу или информационной потребности пользователя [9]. Данная задача может решаться как самостоятельная задача, так и как вспомогательная задача при поиске документов, построении рефератов и поиске ответов на вопросы. Для ее решения разработан широкий набор подходов [1, 8, 9]: на основе вычисления TF-IDF весов у фрагментов, на основе вычисления функции правдоподобия запроса, на основе построения вероятностных моделей для оценки релевантности запроса, на основе использования обучаемых классификаторов, на основе использования скрытой марковской модели, а также на основе комбинировании нескольких подходов.

Задача классификации фрагментов заключается в извлечении фрагментов документа, которые соответствуют одной или нескольким заранее заданным рубрикам [2]. Данная задача несколько отличается от задачи поиска фрагментов, так как предполагает, что рубрики по которым осуществляется набор фрагментов заранее заданы. Для ее решения также разработан набор специализированных методов. Давайте рассмотрим наиболее типичные из них.

В работе [1] авторы предлагают использовать трехэтапный подход к распознаванию фрагментов. На первом этапе производится обучение классификатора с использованием обучающего множества, состоящего из полных текстов. На втором этапе производится разбиение текстов на фрагменты с использованием различных подходов. На третьем этапе осуществляется классификации всех фрагментов с использование классификатора обученного на первом шаге.

В работе [2] авторы используют двух этапную технологию. На первом этапе производится отнесение документа целиком к одной из рубрик с исполь-

зованием простейшего Байесовского классификатора. На втором этапе в тексте документа находится фрагмент, который наиболее соответствует соответствующей рубрике, путем вычисления максимума функции правдоподобия.

В работе [6] предлагается использовать скрытую марковскую модель. Выделение фрагментов происходит в два этапа. На первом этапе для каждой рубрики оцениваются параметры скрытой марковской модели с использованием обучающего множества, состоящего из целых документов. На втором этапе построенная модель используется для выделения фрагментов.

В ряде случаев классификация фрагментов является вспомогательной задачей для повышения качества классификации документов. В работе [10] авторы предлагают выполнять независимую классификацию фрагментов текста, а затем объединять полученные результаты для получения итоговой классификации документов целиком.

Наконец, в работе [7] фрагменты явным образом не выделяются, а используется специальный метод для вычисления весов терминов, который учитывает расположение терминов в тексте. В частности, вектор документа получается как взвешенная сумма векторов предложений.

Необходимо отметить, что во всех рассмотренных выше подходах обучение классификаторов производится на целых документах, т. е. на этапе обучения внутренняя структура документов не учитывается.

В данной работе рассматривается новый итерационный подход к обучению классификаторов, который основан на использовании результатов выделения фрагментов в текстах для итерационного переобучения классификатора. Для этих целей предлагается использовать следующую технологию.

На первом шаге, для каждой рубрики обучается бинарный классификатор с использованием обучающего множества, состоящего из полных документов.

На втором шаге, для каждой рубрики в каждом документе выделяются релевантные фрагменты с использованием построенных классификаторов.

На третьем шаге, для каждой рубрики формируется специальное обучающее множество, которое состоит из фрагментов текстов, выделенных для данной рубрики на предыдущем шаге.

На четвертом шаге, осуществляется проверка критерия завершения работы и либо завершается обучение классификатора, либо повторяются шаги два и три.

Работа имеет следующую структуру. В разделе 2 рассматриваются базовые методы, которые используются для обучения классификаторов. В разделе 3 описываются методы выделения фрагментов. В разделе 5 дается описание новых алгоритмов обучения классификаторов и классификации текстов, с использованием информации о фрагментах. В разделе 5 приводятся результаты экспериментов с различными тестовыми массивами.

2. Используемые методы классификации

Для классификации текстов в настоящей работе используются два базовых метода: метод на основе машин опорных векторов (SVM) и метод на основе байесовского классификатора, в котором в качестве вероятностного распределения для отдельных рубрик используется распределение фон Мизеса-Фишера (VMF). Приведем необходимые определения.

Пусть $\Omega=\{\omega_1,...,\omega_k\}$ — множество из k рубрик и D — множество всех документов. В соответствии с работой [11] задачу классификации текстов будем рассматривать как k независимых задач, т. е. для каждой рубрики ω_j строится решающая функция вида

$$H_j(d) = \begin{cases} 1, & d \in \omega_j, \\ 0, & d \notin \omega_j. \end{cases}$$

Также будем считать, что каждый документ $d \in D$ представляется в виде вектора весов $x=(x_1,...,x_m)^T$, где x_l — вес признака $l=1,...,m$, m — размерность пространства признаков. При проведении экспериментов в данной работе в качестве признаков выступают как отдельные слова, так и словосочетания.

2.1. Метод машин опорных векторов

В методе SVM решающая функция $H_j(d)$, $j=1,...,k$, является линейной функцией следующего вида

$$H_j(d) = H_j(x) = \begin{cases} 1, & R_j(x) > 0, \\ 0, & R_j(x) < 0, \end{cases}$$

где $R_j(x)=w_j^T x+w_{0j}$, $R_j(x)$ — дискриминантная функция, $w_j=(w_{j1},...,w_{jm})^T$ — вектор весов и w_{j0} — смещение, которые определяют гиперплоскость разделяющую документы на классы.

Заметим, что $R_j(x)$ можно рассматривать как меру соответствия документа рубрике ω_j, так как $R_j(x)/\|w_j\|_2$ — расстояние от вектора x до разделяющей гиперплоскости.

Пусть $(x_1,y_1),...,(x_n,y_n)$ — обучающее множество для рубрики ω_j, $j=1,...,k$, где x_i — вектор признаков и $y_i \in \{-1,1\}$ — вектор идентификатор рубрик для документов d_i, $i=1,...,n$. Если документ $d_i \in \omega_j$, тогда $y_i=1$, в противном случае $y_i=-1$.

Параметры разделяющей гиперплоскости w_j и w_{j0} находятся путем максимизации расстояния от разделяющей гиперплоскости положительных и отрицательных примеров, в частности, это делается путем решения следующей оптимизационной задачи

$$\min_{w,w_0} w_j^T w_j + C\sum_{i=1}^{n}\xi_i,$$

$$\begin{cases} y_i\left(w_j^T x_i + w_{j0}\right) \geq 1-\xi_i, i=1,\dots,n, \\ \xi_i \geq 0, i=1,\dots,n, \end{cases}$$

где ξ_i, $i=1,...,n$, вспомогательные переменные, C — положительный параметр.

2.2. Байесовский метод

При использовании байесовского подхода решающая функция H_j для рубрики ω_j, $j=1,...,k$, имеет следующий вид

$$H_j(x) = \begin{cases} 1, & \dfrac{p_j f\left(x|\omega_j\right)}{\overline{p}_j f\left(x|\overline{\omega}_j\right)} > 1, \\ 0, & \dfrac{p_j f\left(x|\omega_j\right)}{\overline{p}_j f\left(x|\overline{\omega}_j\right)} \leq 1, \end{cases}$$

где $p_j=p(\omega_j)$ — вероятность рубрики ω_j, $f(x|\omega_j)$ — функции плотности для рубрики ω_j, $j=1,...,k$ $\overline{\omega}_j$ — обозначает множество документов, которые не относятся к рубрике ω_j, т. е. множество отрицательных примеров.

В данной работе в качестве $f(x \mid \omega_j)$ используется функция плотности распределения фон Мизеса-Фишера, которая определяется следующим образом

$$f\left(x|\omega_j\right) = f\left(x|\mu_j,\kappa_j\right) = c_m\left(\kappa_j\right)e^{\kappa_j \mu_j^T x},$$

где $x \in R^m$, $\|x\|_2=1$, $\mu_j \in R^m$ среднее направление, $\|\mu_j\|_2=1$, $\kappa_j \geq 0$ мера концентрации, $m \geq 2$, $c_m(\kappa)$ нормализующий множитель.

Пусть теперь $(x_1,y_1),...,(x_n,y_n)$ — обучающее множество текстов для рубрики ω_j, $j=1,...,k$, где $x_i \in R^m$ — вектор признаков и $y_i \in \{0,1\}$ — идентификатор рубрики для документа d_i, $i=1,...,n$. При этом, если документ d_i относится к классу ω_j, то $y_i=1$, в противном случае $y_i=0$.

Для построения классификатора $H_j(x)$ для рубрики ω_j, $j=1,...k$, необходимо найти оценки параметров модели фон Мизеса-Фишера p_j, μ_j, κ_j и $\overline{p}_j, \overline{\mu}_j, \overline{\kappa}_j$. В работе [12] показывается, что оценки максимального правдоподобия для данных параметров имеют следующий вид

$$p_j^* = \frac{1}{n}\sum_{i=1}^{n} y_i, \overline{p}_j^* = \frac{1}{n}\sum_{i=1}^{n}(1-y_i),$$

$$\mu_j^* = \frac{r_j^*}{\left\|r_j^*\right\|_2}, \overline{\mu}_j^* = \frac{\overline{r}_j^*}{\left\|\overline{r}_j^*\right\|_2},$$

$$\kappa_j^* \approx \frac{r_j^* m - \left(r_j^*\right)^3}{1-\left(r_j^*\right)^2}, \overline{\kappa}_j^* \approx \frac{\overline{r}_j^* m - \left(\overline{r}_j^*\right)^3}{1-\left(\overline{r}_j^*\right)^2},$$

где $r_j^* = \sum_{i=1}^{n} y_i x_i, \overline{r}_j^* = \sum_{i=1}^{n}(1-y_i)x_i.$

Предварительные эксперименты с различными массивами текстов показали, что в большинстве случаем достаточно использовать фиксированные значения $\kappa_j^*=\overline{\kappa}_j^*=10$. В данном случае решающее правило для рубрики можно упростить и представить в следующем виде

$$H_j(x) = \begin{cases} 1, & R_j(x) > 0, \\ 0, & R_j(x) \leq 0, \end{cases}$$

где $R_j(x) = \kappa_0^{-1} log\frac{p_j}{\overline{p}_j} + \left(\mu_j - \overline{\mu}_j\right)^T x,$

$R_j(x)$—дискриминантная функция, $\kappa_0=10, j=1,...,k$. Заметим, что чем больше значение $R_j(x)$, то тем ближе вектор x центру рубрики ω_j. Следовательно, $R_j(x)$ можно использовать в качестве меры соответствия документа к рубрике ω_j.

3. Выделение фрагментов

Рассмотрим теперь подходы к выделению фрагментов в текстах. Далее будем считать, что фрагмент это набор предложений и что на входе процедуры выделения фрагментов имеется документ, который $d \in D$ представляется с помощью матрицы $X \in R^{(m \times s)}$ и набор независимых классификаторов для каждой рубрики ω_j, $j=1,...,k$, где $X=(x_1,...,x_s)$, $x_t \in R^m$ — вектор весов признаков для предложения $t=1,...,s$, s — число предложений в документе d и m — общее число различных признаков во всех документах.

На выходе процедуры необходимо получить матрицу Λ, которая содержит веса предложений для каждой рубрики ω_j, $j=1,...,k$, где $\Lambda \in R^{(k \times s)}$, $\Lambda=(\lambda_1,...,\lambda_s)$, $\lambda_t=(\lambda_1 t,...,\lambda_k t)^T$, $\lambda_j t$ — вес предложения $t=1,...,s$ для рубрики ω_j, $j=1,...,k$.

Для выделения фрагментов в настоящей работе используются следующие четыре подхода: выделение фрагментов путем классификации предложений, выделении фрагментов путем классификации блоков текста, выделение фрагментов путем классификации иерархического покрытия и выделение фрагментов путем решения специальной оптимизационной задачи. Рассмотрим указанные подходы более подробно.

3.1. Выделение фрагментов путем классификации предложений

Данный подход является наиболее простым и используется во многих работах [1,2,10]. Вес $\lambda_j t$ предложения $t=1,...,s$ для рубрики ω_j, $j=1,...,k$ вычисляется следующим образом

$$\lambda_{jt} = R_j\left(\frac{x_t}{\|x_t\|_2}\right),$$

где $R_j(x)$ — дискриминантная функция для рубрики ω_j. Необходимо отметить, что перед вычислением веса предложения производится нормировка соответствующего вектора признаков таким образом, чтобы его длина была равна 1. Это необходимо из-за того, что обучение классификатора производится с использованием нормированных векторов.

3.2. Выделение фрагментов путем классификации блоков текста

В данном подходе выделение фрагментов производится в два этапа.

На первом этапе текст разбивается на набор смежных блоков предложений с использованием метода TextTiling [3]. При проведении экспериментов использовались следующие параметры данного алгоритма: размер блока равен 4 предложениям, похожесть блоков вычисляется с использованием косинусной меры близости, сглаживание осуществляется с использованием алгоритма скользящего среднего, в котором размер окна равен 3. В результате получается вектор $b=(b_1,...,b_s)$, где $b_t \in \{1,B\}$ номер блока для предложения $t=1,...,s$, B — общее число блоков в тексте.

На втором этапе вычисляются веса предложений $\lambda_j t$, $t=1,...,s$, для рубрик ω_j, $j=1,...,k$ путем использования следующей формулы

$$\lambda_{jt} = u_{jb_t},$$

где u_{ji} — мера соответствия блока $i=1,...,B$ рубрике ω_j, $j=1,...,k$

$$u_{ji} = R_j\left(\frac{\sum_{t\in\{l=1,...,s|b_l=i\}} x_t}{\left\|\sum_{t\in\{l=1,...,s|b_l=i\}} x_t\right\|_2}\right).$$

Необходимо отметить, что также можно использовать и другие методы разделения текстов на блоки [3,4,5,6].

3.3. Выделение фрагментов путем классификации иерархического покрытия

В данном подходе вес предложения $\lambda_j t$, $t=1,...,s$, для рубрики ω_j, $j=1,...,k$, вычисляется путем нахождения суммы весов всех непрерывных фрагментов содержащих данное предложение. В частности, $\lambda_j t$ находится следующим образом

$$\lambda_{jt} = \sum_{y\in F, x_t\in y} R_j\left(\frac{y}{\|y\|_2}\right),$$

где F множество векторов признаков, соответствующих всем непрерывным фрагментам предложений в данном документе, т. е.

$$F = \left\{\sum_{t=l_1}^{l_2} x_t \;|1 \le l_1 \le l_2 \le s\right\}.$$

Заметим, что для вычисления весов для всех предложений для одной категории необходимо вычислить степень соответствия всех фрагментов данной рубрике, что требуется порядка $O(s^2)$ операций.

Для снижения вычислительной сложности можно воспользоваться иерархическим множеством фрагментов H вместо использования всего множества фрагментов F, где

$$H = H_0 \cup \left(\bigcup_{i=1}^{\lceil \log_2 s\rceil} H_i\right),$$

$$H_t = \left\{\sum_{t=1+l2^{i-1}}^{\min(l2^{i-1}+2^i,s)} x_t \;|l = 0,\dots,\left\lceil\frac{s}{2^{i-1}} - 1\right\rceil\right\},$$

$$H_0 = \{x_1,....,x_s\}.$$

Можно показать, что

$$|H| = |H_0| + \sum_{i=1}^{\lceil \log_2 s\rceil} |H_i| \le \log_2 s + 1 + 3s.$$

Таким образом, в данном случае вычислительная сложность порядка $O(s)$. Также можно показать, что для любого фрагмента $f \in F$ существует фрагмент $h \in H$ такой, что

$$\frac{|h\Delta f|}{|f|} \le \frac{1}{2},$$

где $|f|$ — число предложений во фрагменте f, $|f\Delta h|$ число различных предложений во фрагментах f и h. Иными словами, для любого фрагмента $f \in F$ существует фрагмент $h \in H$, который содержит, по крайней мере, половину общих предложений. В результате, выражение для вычисления весов предложений приобретает следующий вид

$$\lambda_{jt} = \sum_{h\in H, x_t\in h} R_j\left(\frac{h}{\|h\|_2}\right).$$

3.4. Выделения фрагментов с использованием оптимизационных методов

В данном новом подходе веса $\lambda_j=(\lambda_j1,...\lambda_js)$ предложений для рубрики ω_j, $j=1,...,k$ находятся путем решения следующей оптимизационной задачи

$$\max_{\lambda_j} R_j\left(z(\lambda_j)\right),$$

где $z(\lambda_j)$ — вектор документа, основанный на векторах предложений, который определяется следующим образом

$$z(\lambda_j) = \frac{\sum_{t=1}^{s} \lambda_{jt} x_t}{\left\|\sum_{t=1}^{s} \lambda_{jt} x_t\right\|_2}.$$

Заметим, что из определения следует, что $\|z(\lambda_j)\|_2=1$.

Рассмотрим теперь процедуру нахождения приближенного решения данной задачи для случая, когда в качестве решающей функции используется классификатор на основе машин опорных векторов. В данном случае оптимизационная задача имеет следующий вид

$$\max_{\lambda_j} w_j^T z(\lambda_j) + w_{0j}.$$

Прямой нахождение решения данной оптимизационной задачи является достаточно сложным. По этой причине рассмотрим следующую упрощенную задачу

$$\max_{\|\xi\|_2=1} w_j^T \xi + w_{0j}.$$

Несложно показать, что в данном случае максимум достигается, когда $\xi=w_j/\|w_j\|_2$.

Таким образом, можно найти приближенное решение исходной задачи путем нахождения решения следующего уравнения

$$z(\lambda_j)=\xi,$$

которое можно переписать следующим образом

$$\frac{\sum_{t=1}^{s} \lambda_{jt} x_t}{\left\|\sum_{t=1}^{s} \lambda_{jt} x_t\right\|_2} = \frac{w_j}{\left\|w_j\right\|_2}.$$

Последнее равенство эквивалентно следующему равенству

$$\sum_{t=1}^{s} \lambda_{jt} x_t = w_j.$$

В матричной форме оно имеет следующий вид

$$X\lambda_j^T=w_j.$$

Необходимо отметить, что матрица X имеет m строк и s столбцов, где m — число различных признаков и s число предложений. В большинстве случаев $m \gg s$. Более того, в большинстве случаев матрица X является сильно разряженной. По этой причине приведенное выше равенство обычно не имеет точного решения и для нахождения решения необходимо пользоваться приближенными методами, ориентированными на работу с большими разряженными матрицами.

В данной работе для этих целей используется алгоритм LSQR [13], который как раз ориентирован для работы с такими данными. При его использовании решение уравнения находится путем итерационного решения задачи следующей задачи наименьших квадратов

$$\min_{\lambda_j} \left\|X\lambda_j^T - w_j\right\|_2^2.$$

Пусть $\lambda_j^*=(\lambda_{j1}^*,...,\lambda_{js}^*)$ соответствующее решение последней задачи. Тогда, если $\lambda_{jt}^*>0$ можно считать, что предложение $t=1,...,s$ соответствует «положительной» рубрике ω_j. По этой причине в качестве предложений соответствующих рубрике можно рассматривать те, которые соответствуют положительным элементам вектора λ_j^*.

Рассмотрим теперь выделение значимых предложений в том случае, когда используется байесовский классификатор на основе распределения фон Мизеса-Фишера. В данном случае исходная оптимизационная задача имеет следующий вид

$$\max_{\lambda_j} \left(\mu_j - \overline{\mu}_j\right)^T z(\lambda_j) + \kappa_0^{-1} log \frac{p_j}{\overline{p}_j}.$$

Если обозначить $w_j^T = \left(\mu_j - \overline{\mu}_j\right)^T$ и

$w_{j0} = \kappa_0^{-1} log \frac{p_j}{\overline{p}_j}$, тогда исходная оптимизационная задача будет полностью совпадать с исходной оптимизационной задачей для SVM классификатора. Следовательно, решение ее можно находить аналогичным образом.

4. Обучение классификаторов

Рассмотрим теперь алгоритмы обучения классификаторов и классификации документов на основе фрагментов.

4.1. Обучение классификаторов с использованием фрагментов

Пусть $\Omega=\{\omega_1,...,\omega_k\}$ множество из k заранее заданных рубрик и $d_1,...,d_n$ обучающее множество документов. Каждый документ d_i, $i=1,...,n$, в обучающем множестве представляется с помощью пары $\langle X_i,c_i\rangle$, где $X_i=(x_{1i},...,x_{(si\,i)})$ — матрица, столбцы которой являются векторами весов информационных признаков для предложений документа, s_i — число предложений в документе и $c_i=(c_{1i},...,c_{ki})^T$ вектор с идентификаторами рубрик, причем $c_{ji}\in\{0,1\}$ признак отнесения документа d_i к рубрике ω_j экспертом.

Пусть $\Lambda_i=(\lambda_{jti})_{k\times si}$, $i=1,...,n$ множество матриц таких, что λ_{jti} вес предложения $t=1,...,s_i$ для рубрики $j=1,...,k$ в документе с номером $i=1,...,n$. Тогда алгоритм обучения классификатора принимает следующий вид.

1. Положить $iter=1, \lambda_{jti}=1$ for $i=1,...,n, j=1,...,k$, $t=1,...,s_i$.
2. Вычислить вектора информационных признаков z_{ij} для каждого документа d_i, $i=1,...,n$, и рубрики $\omega_j, j=1,...,k$, путем использования следующей формулы

$$z_{ij}=\frac{\sum_{t=1}^{s_i}\lambda_{jti}x_{ti}}{\left\|\sum_{t=1}^{s_i}\lambda_{jti}x_{ti}\right\|_2}$$

3. Построить для каждой рубрики $\omega_j, j=1,...,k$, бинарный классификатор H_j путем использования векторов документов $z_{1j},...,z_{nj}$ и эталонных индикаторов принадлежности документов к данной рубрике $c_{j1},...,c_{jn}$.
4. Для каждой рубрики $\omega_j, j=1,...,k$, выделить фрагменты во всех документах и перевычислить значения весов Λ_i, $i=1,...,n$ путем использования одного из методов выделения фрагментов и классификаторов H_j, обученных на предыдущем шаге.
5. Если $iter$ меньше заданного порога, то перейти к шагу 2, в противном случае завершить работу алгоритма.

4.2. Классификация текстов с использованием фрагментов

Теперь давайте рассмотрим алгоритм классификации документов, который применяется совместно с описанным алгоритмом обучения классификаторов.

Пусть имеется документ $d\in D$, который представлен с помощью матрицы $X\in R^{m\times s}$ и классификаторы H_j для каждой рубрики ω_j, $j=1,...,k$, где $X=(x_1,...,x_s)$, $x_t\in R^m$ — вектор весов информационных признаков для предложения $t=1,...,s$, s число предложений в документе d и m число различных информационных признаков во всех документах D.

Пусть $\Lambda\in R^{k\times s}$ обозначает матрицу весов предложений для рубрик ω_j, $j=1,...,k$, где λ_{jt} вес предложения $t=1,...,s$ для рубрики ω_j, $j=1,...,k$. Тогда алгоритм классификации текстов имеет следующий вид.

1. Для каждой рубрики $\omega_j, j=1,...,k$, выделить соответствующие ей фрагменты в документе и вычислить веса Λ путем использования одного из подходов к выделению фрагментов и классификатора H_j.
2. Для каждой рубрики ω_j, $j=1,...,k$, вычислить вектора весов документов $z_j\in R^m$ путем использования следующего выражения

$$z_j=\frac{\sum_{t=1}^{s}\lambda_{jt}x_t}{\left\|\sum_{t=1}^{s}\lambda_{jt}x_t\right\|_2}.$$

3. Для каждой рубрики ω_j, $j=1,...,k$, выполнить классификацию векторов документов z_j путем использования классификатора H_j.

Таким образом, в приведенном алгоритме классификации для каждой рубрики строится специальное представление документов, что позволяет адаптироваться к особенностям каждой категории при классификации документов.

5. Эксперименты

5.1. Тестовые массивы текстов

Рассмотрим теперь результаты предварительных экспериментов по оценке качества разработанных методов классификации документов.

При проведении экспериментов использовались два английских массива текстов “Reuters-21578” (http://www.daviddlewis.com/resources/testcollections/reuters21578/) и “20 News Groups” (http://kdd.ics.uci.edu/databases/20newsgroups/20newsgroups.html), и один русский массив текстов “ROMIP-2004” (http://romip.ru/en/index.html). В Таблице 1 приведены основные характеристики данных массивов.

Таблица 1. Основные характеристики массивов текстов

Массив	Число документов	Число рубрик	Размер массива
Reuters-21578	13476	123	16.2 Mb
20 News Groups	16330	20	46 Mb
ROMIP 2004	6931	173	281 Mb

Также при проведении экспериментов использовались сокращенные варианты данных

массивов: “Reuters-21578-10”, “20 News Groups Mini”, “ROMIP 2004 Mini”. Данные сокращенные массивы более подходят для интенсивных экспериментов ввиду меньшего размера и более сбалансированной структуры. Более того данные массивы значительно более часто используются в работах других исследователей, чем исходные полные массивы [1,2].

Массив “Reuters-21578-10” состоит из десяти наибольших рубрик “acq”, “corn”, “crude”, “earn”, “grain”, “interset”, “money-fx”, “ship”, “trade”, “wheat” из массива “Reuters-21578”.

Массив “20 News Groups Mini” является подмножеством массива “20 News Groups” и содержит 2000 случайным образом отобранных документов. В каждой рубрике при этом ровно по 100 документов.

Массив “ROMIP-2004 Mini” является подмножеством массива “ROMIP 2004”. Он включает десять наибольших рубрик со следующими номерами: 9001149, 9001227, 9001355, 9001423, 9001557, 9001613, 9001755, 9001759, 9001716900, 901800651. В таблице 2 приведены основные характеристики данных массивов.

Таблица 2. Основные характеристики сокращенных тестовых массивов текстов

Массив	Число документов	Число рубрик	Размер
Reuters-21578-10	8592	10	8.9 Mb
20 News Groups Mini	1954	20	4.7 Mb
ROMIP 2004 Mini	1704	10	113 Мб

Для оценки качества работы классификаторов в настоящей работе использовались стандартные показатели: точность, полнота, F-мера. При вычислении данных показателей для всего массива в целом использовалось макроусреднение. Вычисление значений показателей производилось в соответствии со следующей схемой.

1. Разбиение эталонного массива текстов на обучающее и тестовое множество случайным образом в заданной пропорции.
2. Обучение классификатора на обучающем множестве с использованием различных методов.
3. Классификация документов из тестового множества с использованием построенных классификаторов и оценка качества классификации.

Необходимо отметить, что все алгоритмы классификации и обработки текстов были реализованы на языке MATLAB в виде специальных функций и классов.

Далее при описании экспериментов будут использоваться следующие сокращенные обозначения: SVM — стандартный классификатор SVM, который обучен на полных документах; VMF — стандартный байесовский классификатор на основе распределения фон Мизеса-Фишера, который обучен на полных документах; SVM-SENT — классификатор SVM, обученный на фрагментах, выделенных с помощью классификации предложений; SVM-TILE — классификатор SVM, обученный на фрагментах, выделенных путем классификации блоков предложений; SVM-HIER — классификатор SVM, обученный на фрагментах, выделенных путем классификации иерархического покрытия документа; SVM-LS — классификатор SVM, обученный на фрагментах, выделенных путем решения специальной оптимизационной задачи; VMF-SENT — классификатор VMF, обученный на фрагментах, выделенных с помощью классификации предложений; VMF-TILE — классификатор VMF, обученный на фрагментах, выделенных путем классификации блоков предложений; VMF-HIER — классификатор VMF, обученный на фрагментах, выделенных путем классификации иерархического покрытия документа; VMF-LS — классификатор , обученный на фрагментах, выделенных путем решения специальной оптимизационной задачи.

5.2. Результаты экспериментов

Рассмотрим теперь результаты предварительных экспериментов с сокращенными массивами текстов. В данных экспериментах исходное эталонное множество документов разделялось на обучающее и тестовое множества в пропорции 75% на 25%. В таблицах 3 и 4 приведены показатели качества классификации для методов SVM и SVM-LS для массивов “20 News Groups Mini” и “ROMIP 2004 Mini”, соответственно.

Из приведенных таблиц можно заметить, что качество работы метода SVM-LS значительно выше качества работы стандартного метода SVM.

Анализ результатов обработки массива 20 News Groups показывает, что повышение качества классификации возможно связано с тем, что оригинальные документы содержат служебные заголовки почтовых сообщений, которые являются малоинформативными, а также рассуждения участников новостных групп на отвлеченные темы, которые не относятся напрямую к основной теме новостной группы.

Анализ результатов обработки массива ROMIP 2004 Mini показывает, что повышение качества классификации, скорее всего, связано с политематическим содержанием нормативно-правовых документов. Также необходимо отметить, что достигнутые показатели качества классификации не уступают тем, которые получены другими авторами в рамках семинара РОМИП (в данных работах F-мера принимает значения от 0,1 до 0,5 в зависимости от состава используемых рубрик).

Таблица 3. Качество работы классификаторов для массива 20 News Groups Mini

Method	Precision	Recall	F-measure
SVM	0.94	0.27	0.40
SVM-LS	0.96	0.89	0.92

Таблица 4. Качество работы классификаторов для массива ROMIP 2004 Mini

Method	Precision	Recall	F-measure
SVM	0.67	0.29	0.36
SVM-LS	0.76	0.39	0.50

Анализ качества работы всех рассмотренных алгоритмов на массиве ROMIP 2004 Mini приведен в таблице 5, из которой можно заметить, что в наибольшей степени повышение качества классификации происходит при использовании алгоритма SVM.

Таблица 5. Качество работы классификаторов для массива ROMIP 2004 Mini

Method	F-measure	Method	F-measure
SVM	0.37	VMF	0.45
SVM-SENT	0.39	VMF-SENT	0.47
SVM-HIER	0.38	VMF-HIER	0.46
SVM-TILE	0.39	VMF-TILE	0.46
SVM-LS	0.50	VMF-LS	0.37

Предварительные эксперименты с другими массивами также показали, что повышение качества классификации в основном происходит при относительно небольшом размере обучающего множества. Для проверки данной гипотезы была проведена серия экспериментов, в которых в качестве параметра выступал размер обучающего множества. В частности, на рисунках 1 и 2 приведены зависимости качества классификации от размера обучающего множества для различных алгоритмов на массивах “20 News Groups” и “Reuters-21578”.

Проведенные эксперименты подтвердили гипотезу о том, что качество классификации более сильно увеличивается при маленьких объемах обучающих выборок. Также было установлено, что наиболее эффективным является метод SVM-LS.

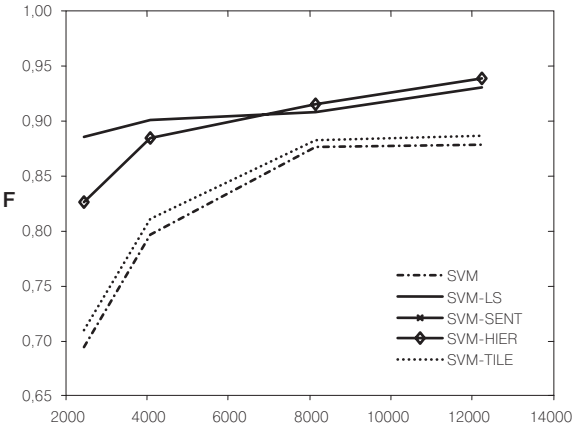


Рис. 1. Качество работы SVM классификаторов на массиве 20NG

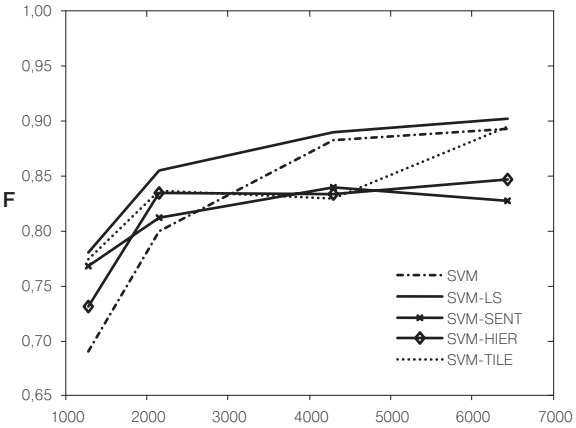


Рис. 2. Качество работы SVM классификаторов на массиве Reuters-21578-10

6. ВЫВОДЫ

Таким образом, в настоящей работе предложен новый подход к обучению классификаторов, основанный на использовании результатов автоматического выделения фрагментов в текстах. Экспериментально показано, что в ряде случаев данный подход может быть достаточно эффективным значительно повышать качество классификации. Также в работе рассмотрен новый метод выделения фрагментов, основанный на решении специальной оптимизационной задачи и нахождении весов предложений в тексте.

Литература

1. *Mengle S., Goharian N.* Passage detection Using Text Classification. Journal of the American Society for Information Science and Technology, 60(4), 2009, pp. 814–825.
2. *Murtagh Y. B., McClean S., Anderson T.* Text Passage classification using supervised Learning, 1999. — 13 p. (http://citeseerx.ist.psu.edu/viewdoc/summary?doi=?doi=10.1.1.130.4741)
3. *Hearst M.* TextTiling: Segmenting Text into Multi-Paragraph Subtopic Passages. Computational Linguistics , 23 (1), 1997. pp. 33–64.
4. *Choi F., Wiemer-Hasting P., Moore J.* Latent semantic Analysis for Text Segmentation. Proceedings of NAACL'01, Pittsburgh, PA, 2001. — pp. 109–117.
5. *Blei D., Moreno P. J.* Topic Segmentation with an Aspect Hidden Markov Model. SIGIR'01, September 9–12, 2001, New Orleans, Louisiana, USA. — 6 p.
6. *Denoyer L., Zaragoza H., Gallinary P.* HMM-based Passage Models for Document Classification and Ranking. 23-s BCS European Annual Colloquium on Information Retrieval, 2001.
7. *Ko Y., Park J., Seo J.* Improving text categorization using the importance of sentences. Information Processing and Management 40, 2004. — pp. 65–79.
8. *Extraction* of Coherent Relevant Passages Using Hidden Markov Model. ACM Transactions on Information Systems, Vol. 24, No. 3, 2006. pp. 295–319.
9. *Wade C., Allan J.* Passage Retrieval and Evaluation. CIIR Technical Report IR-396, 2005. — 9p. (http://ciir.cs.umas.edu/pubfiles/ir-396.pdf)
10. *Kim J., Kim M.* An evaluation of passage-based text categorization. Journal of Intelligent Information Systems, 23, 2004. pp. 47–65.
11. *Sebastiani F.* Machine learning in automated text categorization. ACM Computing Surveys, 34(1), 2002. — pp. 1–47.
12. *Text* Mining: classification, clustering, and applications / Srivastava A., Sahami M. CRC Press, 2009. — 290 p.
13. *Paige C. C., Saunders M. A.* LSQR: An Algorithm for Sparse Linear Equations And Sparse Least Squares. ACM Trans. Math. Soft., Vol.8, 1982. — pp.43–71.

Топологический аспект многозначности идиом

Topological aspect of polysemy in semantics of idioms

Вознесенская М. М. (voznes-masha@yandex.ru)

Институт русского языка им. В. В. Виноградова РАН

Работа посвящена типам иерархической организации многозначности в идиоматике. Рассматриваются радиальный, цепочечный и смешанный (радиально-цепочечный) топологические типы многозначности идиом. Предлагается способ определения типа многозначности идиом с двумя значениями.

I. Лексическая многозначность в русском языке уже достаточно хорошо описана [Апресян 1974; Кустова 2004; Падучева 2004; Розина 2005, Зализняк Анна 2006] (предварительные итоги см., например, в [Апресян 2009: 409–485]) Однако во фразеологии изучение многозначности только начинается [Баранов, Добровольский 2008: 454–472]. Между тем исследование различных аспектов структуры полисемии идиом — одна из важнейших задач как семантического описания фразеологии, так и семантики в целом. В этой работе мы рассмотрим многозначность идиом с точки зрения иерархических отношений между значениями. Материалом исследования послужили многозначные идиомы из «Фразеологического объяснительного словаря русского языка» (далее Словарь) [Баранов, Вознесенская, Добровольский и др. 2009]. Из более 1000 идиом, содержащихся в Словаре, 321 — многозначна. При этом 227 идиом имеет два значения, у 70 идиом выделяется по три значения, у 24 — более трех значений. Максимальное количество, шесть значений, имеют 2 идиомы: *оставить мокрое место* (с вариантом *мокрого места не оставить*) и *поставить/поднять… на ноги*.

II. В лексической семантике принято выделять три топологических типа многозначности: радиальную, цепочечную и радиально-цепочечную. При радиальной полисемии «все значения слова мотивированы одним и тем же — центральным — значением», в случае же цепочечной полисемии "каждое новое значение слова мотивировано другим — ближайшим к нему — значением, но крайние значения могут и не иметь общих семантических компонентов" [Апресян 1974: 182]. При радиально-цепочечном типе в структуре многозначного слова представлены оба топологических типа полисемии. Возможно ли выделение этих типов многозначности в сфере идиоматики? Вопрос не так прост, как может показаться вначале. Дело в том, что разграничение радиальной и цепочечной полисемии предполагает наличие у слова не менее трех значений, иначе противопоставление между этими типами многозначности нейтрализуется. Большинство же идиом имеет только два значения и, казалось бы, не попадают под топологическую классификацию. Однако при сравнении смысловых структур таких идиом с двумя значениями, как *из-под полы*, с одной стороны, и *ползать на брюхе* (перед кем-л.); *приползти на брюхе*, с другой, мы увидим несомненное различие.

ИЗ-ПОД *ПОЛЫ*[1] 1. В нарушение законов продавать в не предназначенных для этого местах какие-л. товары или предлагать какие-л. услуги так, чтобы проверяющие инстанции этого не обнаружили, *что описывается как сокрытие предлагаемого товара под одеждой.* Чтобы извлечь из сумки журнал, Маше пришлось сначала вывалить на стол две пачки пельменей, потом немецкий маргарин, а найдя «бесценное издание», она отпихнула его на край стола, стараясь не зацепиться взглядом за эротическое откровение на обложке. — Торгуют-то этим не *из-под полы*! — осуждающе проговорила Костикова. —

[1] Здесь и далее словарные статьи идиом взяты из «Фразеологического объяснительного словаря русского языка» [Баранов, Вознесенская, Добровольский и др. 2009]. В ряде случаев количество примеров, иллюстрирующих употребление идиом в разных значениях, сокращается для экономии места.

Кстати, знаете где? В двух шагах от Российской Академии образования! И никто, между прочим, там в обморок не упал. Ю. Каменецкий и др. Мелочи жизни.
2. Делать что-л. тайно, не имея официального разрешения, *что описывается по аналогии с ситуацией незаконной торговли.* 🗐 Работа была, может быть, не строго научна, но, пожалуй, талантлива и написана хорошо по-русски, таким летящим, взмывающим слогом, но главное, что и поразило, что и произвело… была внутренне свободна. Мы видели ее однажды на кафедре, уже желтую, с потрепанными ушами… Она там хранилась, по-видимому, как беспрецедентный случай. Ею гордились не перечитывая, и кое-кому, *из-под полы*, показывали. А. Битов. Пушкинский дом.

В идиоме *из-под полы* первое значение ‘незаконная торговля’ мотивировано внутренней формой, в которой описывается реальный способ совершения таких сделок. Второе значение производно от первого, когда ‘незаконная торговля’ переосмысляется как ‘совершение чего-л. тайного, не имеющего официального разрешения’. Можно рассматривать подобную структуру многозначности как цепочечную, в которой внутренняя форма идиомы, обозначающая реальную ситуацию, предстает как «исходное значение», которым мотивировано первое значение, мотивирующее, в свою очередь, последующее, второе, значение.

Структура многозначности идиомы *ползать на брюхе* (перед кем-л.); *приползти на брюхе* устроена по-другому:

ПОЛЗАТЬ НА *БРЮХЕ* (перед кем-л.) *снижен.*; **ПРИПОЛЗТИ НА БРЮХЕ** *снижен.* **1**. Очень сильно и униженно просить кого-л. о чем-л., *что осмысляется как передвижение по земле в горизонтальном положении в ногах другого человека (после значимого действия — перехода из вертикального положения), показывающее подчинение просящего и подчеркивающее более низкий социальный статус просящего или снижающий его.* ❖ просить кого-л., умолять кого-л. 🗐 Свободный человек на свободной земле. Захотел на Мадагаскар — поехал на Мадагаскар… — Десять тысяч туда и обратно! — Умолкни! Слышать не хочу. Ты знаешь, сколько я унижался, *на брюхе* перед гадами *ползал*, чтоб в эти Канны попасть? В. Некрасов. Маленькая печальная повесть • Ничего не спасет театр, если директор не будет *ползать на брюхе* перед богатыми спонсорами и вымаливать деньги. Корпус Публ. **2**. Осознавая превосходство другого, выражать готовность признавать это самоуничижением, *что осмысляется как передвижение по земле в горизонтальном положении в ногах другого человека (после значимого действия — перехода из вертикального положения), показывающее подчинение первого человека и подчеркивающее его более низкий социальный статус или снижающий его.* ❖ восхищаться кем-л., превозносить кого-л., боготворить кого-л. 🗐 На вечере воспоминаний было рассказано немало забавных историй. Чего стоит, например, свидетельство Юрия Кублановского: «Юрий Иваск рассказывал мне, что однажды он сказал Бродскому: “Иосиф, я стою перед Державиным на коленях”, на что Бродский ответил: “А я *ползаю* перед ним *на брюхе*”». Корпус Публ. • Если коммунистическая оппозиция разобщена, то демократов, увы, и с фонарем не найдешь — разобщаться некому. Президент верно сказал-сплюнул: «Куда ж они денутся?» Никуда и не делись, как выражался другой политик, «*на брюхе приползли*». Корпус Публ.

Внутренняя форма этой идиомы указывает на конкретное физическое действие человека, занимающего определенное положение в пространстве. Эта реальная ситуация по-разному переосмысляется в каждом из значений идиомы. В первом значении это интерпретируется как ‘униженная просьба’, во втором — как ‘самоуничижительное признание превосходства другого человека’. Оба значения непосредственно связаны с «исходным значением» — внутренней формой, т. е. это радиальный тип многозначности.

Необходимо отметить, что принятая в Словаре форма толкования, включающая не только актуальное значение идиомы, но и ее образную составляющую — внутреннюю форму (выделяемую в тексте толкования курсивом), в ряде случаев содержит прямое указание на «направление» мотивации, и тем самым на тип многозначности.[2]

Таким образом, если рассматривать внутреннюю форму идиомы в качестве исходного мотивирующего значения всех последующих значений, становится возможным определить семантическую структуру идиом с двумя значениями.[3] Тем самым в сфере идиоматики, где идиомообразование представляет собой «особый тип вторичной номинации» [Телия 1996: 134], существенно увеличивается количество единиц (идиом), многозначность которых может быть охарактеризована в топологическом аспекте. Рассмотрим каждый вид топологической многозначности идиом подробнее.

III. Радиальная полисемия. Этот тип многозначности, по данным Словаря, является наиболее распространенным. В нем представлены идиомы с различным количеством значений — от двух до шести. Значения могут быть мотивированы или каким-либо основным значением, являющимся образной переинтерпретацией внутренней формы, либо самой внутренней формой непосредственно.

---

[2] Подробнее об отражении внутренней формы в словарном толковании идиомы см. во Вступительной статье Словаря [Баранов, Вознесенская, Добровольский и др. 2009].

[3] Естественно, что речь идет только об идиомах с такой живой внутренней формой, когда основание мотивации понятно для носителя языка.

В первом случае количество значений должно быть не меньше трех, во втором возможны и два значения (ср. выше *ползать на брюхе* (перед кем-л.); *приползти на брюхе*).

В идиоме *двойная бухгалтерия* первое значение ‘такой способ организации финансовой отчетности, вводящий в заблуждение, при котором создаются два вида документов — настоящие и ложные, скрывающие реальные расходы и доходы и предназначенные для официального отчета, налоговых органов и т. д.’ является основным, которое мотивирует два других: ‘такой способ организации отчетности, *осмысляемый как вводящий в заблуждение способ организации финансовой отчетности,* при котором создаются два вида документов — настоящие и ложные, скрывающие реальное положение дел и предназначенные для обнародования’ и ‘такой способ мышления, при котором человек легко меняет свои принципы, рассуждая о разных людях, событиях, ситуациях и т. д., *что осмысляется как вводящий в заблуждение способ организации финансовой отчетности.*’ Направление мотивации отражается и в толкованиях: внутренняя форма второго и третьего значений является первым актуальным значением. Ср. контексты употреблений на каждое из значений:

‘двойная бухгалтерия 1’:
(1) «Боже мой! ***Двойная бухгалтерия***! Все эти шахер-махер, которые так ловко проделывала бывшая начальница Фаина Ивановна, проделывает Лена», — испугалась Вера Александровна. — Леночка, так надо срочно оттуда уходить! Переедете в Москву, уж куда-куда, а в бухгалтерию я вас определенно устрою! (Л. Улицкая)
‘двойная бухгалтерия 2’:
(2) — …Как вы понимаете, я не щажу и себя. Будучи аспиранткой Ивана Ильича Стригалева, видя все это, видя ***двойную бухгалтерию***, которую вел мой руководитель… А он уже год назад чувствовал, что идут черные для вейсманизма-морганизма времена, и завел два журнала. Два! (В. Дудинцев)
Как видно из примера, второе значение не относится к деятельности в сфере финансов.
‘двойная бухгалтерия 3’:
(3) На нынешнюю власть надо давить, но так, чтобы не задавить. Кто-то скажет — это ***двойная бухгалтерия***. А разве вся наша жизнь не ***двойная бухгалтерия***: знаем прекрасно, что умрем, а стараемся жить и быть счастливыми? Компромисс не порок, а условие жизни. (Известия)

В идиоме *как из пушки* нет эксплицитно выраженного основного значения. Здесь исходным «пунктом» служит обозначенная внутренней формой реальная ситуация (выстрел из пушки), разные признаки (звук, скорость) и следствия (точность попадания) которой образно переосмысляются в каждом из значений:

КАК ИЗ *ПУШКИ* 1. (о звуке) громко, резко и неожиданно, *что сравнивается со звуком выстрела орудия* 🗐 Шум моторов утих, только слышно — ветер свистит. Потом хлопнуло где-то вверху, *как из пушки*, кабина вздрогнула, рванулась и тихо стала приземляться. Корпус Детект.
2. с большой интенсивностью или скоростью, *что сравнивается со скоростью выстрела из орудия и его силой* ❖ быстро, сильно, резко 🗐 — Жить будем, лейтенант, или окоченевать, как цуцики? Жрать хочется, — *как из пушки*! Умираю от голода. Что затихли, заснули все? Ты чего умолк, лейтенант? Ю. Бондарев. Горячий снег.
3. так, что всегда реализуется один вариант развития событий, *что сравнивается с неизбежностью последствий орудийного выстрела* ❖ обязательно, наверняка, непременно 🗐 Я еще два раза попробовал, для опыта. *Как из пушки* маслом книзу, и все! — А потом? — спросил Супонин. — Потом съел, — сказал Брикетов. М. Мишин. Почувствуйте разницу.

IV. Цепочечная полисемия. Этот тип многозначности не так широко представлен, как радиальный, что сближает фразеологию с лексикой, где цепочечная полисемия «в чистом виде редка» [Апресян 1974: 182]. Длина «цепочек» тоже невелика: среди многозначных идиом Словаря цепочечной структурой обладают только идиомы с двумя значениями (ср. выше *из-под полы*). Приведем еще один пример идиомы с цепочечной полисемией:

ПОД *КАБЛУКОМ* 1. (у кого-л./чьим-л.) (о мужчине) (Быть, находиться) в зависимости от жены (или другого лица женского пола, с которым, как правило, связан сексуальными или родственными отношениями), выполняя все ее желания и прихоти, а также разделяя (по крайней мере, в основном) ее отношение к действительности, *что описывается как положение зависимого субъекта под характерной частью женской обуви — вариант реализации жеста, при котором человек наступает на другого человека, выражая свою полную власть над ним*; **ДЕРЖАТЬ ПОД КАБЛУКОМ (кого-л.); ПОПАСТЬ ПОД КАБЛУК (к кому-л.).** 🗐 Витька полгода назад женился и сразу *попал под каблук* к молодой супруге. Характер у него был мягкий, и скорей всего страх перед женой окажется сильней требований Мамонта о строгой конспирации — рано или поздно он расскажет ей все. Корпус Детект.
2. (у кого-л./чьим-л./у чего-л.) (быть, находиться) в зависимости от кого-л./чего-л., *что осмысляется как положение мужа, находящегося в зависимости от своей жены и выполняющего все ее желания и прихоти;* **ДЕРЖАТЬ ПОД КАБЛУКОМ (кого-л./что-л.); ПОПАСТЬ ПОД КАБЛУК (к кому-л./к чему-л.)** 🗐 <…> чтобы заставить крупный бизнес работать на общество, нужно в первую очередь реформировать бюрократию, которая заинтересована в том,

чтобы держать бизнес в тени, а значит — выталкивать его в нелегальное, криминальное пространство. Ведь «олигарха», на которого в прокуратуре есть досье, легче *держать под каблуком* и использовать как денежный мешок. Корпус Публ.

Словарные толкования значений хорошо демонстрируют цепочечный характер многозначности этой идиомы. Так, первое значение ‘зависимости мужа от жены’ основано на образном переосмыслении реальной ситуации ‘нахождения под каблуком’, описание которой включается в первое толкование в качестве внутренней формы для актуального значения. Второе значение ‘зависимости от кого-л.’ мотивировано первым актуальным значением, которое входит во второе толкование уже как образное основание, внутренняя форма второго актуального значения.

V. Смешанная (радиально-цепочечная) полисемия. Этот тип полисемии, в отличие от лексики, для которой он наиболее типичен, среди рассмотренных нами примеров встречается достаточно редко. Проиллюстрируем смешанный тип полисемии идиомой с максимальным набором значений, отразив в нумерации и порядке расположений значений структуру ее многозначности[4]:

ПОСТАВИТЬ/ПОДНЯТЬ… НА *НОГИ* […н'анъги] **1. (кого-л.)** сделать кого-л. здоровым, *что описывается как придание человеку устойчивого вертикального положения, рассматриваемого как норма* ❖ вылечить кого-л. 🗐 — Аку, аку-пунктура! — завопил консул. — Только иглоукалывание спасет нашего друга. Берусь за умеренное вознаграждение *поставить* академика *на ноги*. Я лучший знаток тибетской медицины в Юго-Восточной Азии, и если бы не интриги… — Давайте вернемся к этому позднее, — сказала бабушка. — Не вертитесь, пожалуйста, господин генеральный консул. В. Аксенов. Мой дедушка — памятник.

1.1. (что-л.) сделать так, чтобы что-л., существовавшее или функционировавшее ранее, опять стало существовать или функционировать, *что осмысляется как излечение и выздоровление человека* ❖ восстановить что-л., возродить что-л., укрепить что-л. 🗐 Теперь он раскаялся. Ценя свои сочинения очень высоко, он был уверен, что советская власть рухнула исключительно благодаря им. Но увидев, какие силы теперь пришли к власти, он устыдился прежних своих книг, высказываний и действий, жалел, что разрушил советский строй, каялся и обещал новыми сочинениями этот строй *поставить* обратно *на ноги*. В. Войнович. Монументальная пропаганда

2. (кого-л.) вырастить и сделать самостоятельным в жизни (обычно о воспитании детей), *что описывается как придание человеку устойчивого вертикального положения, рассматриваемого как норма* ❖ воспитать кого-л. 🗐 А обычная мать, он сказал, у меня умерла, может быть, и отца постоянного я лишен в результате алкоголизма, слышал только, величали Ильей. Яша, милый, да может, я он и есть, небось, случались ребятишки какие-нибудь впопыхах, жизнь же тоже огромна. Допускаю, Ильич отвечал, но зачем ты в подобном случае мать забросил с концами, сына женщине *поставить на ноги* не помог, образования ремесленного ему не дал, подлец ты мне после этого, а не отец. И обиделся. Яков Ильич, я утешил, да ты не серчай, я еще, может, и не отец тебе никакой, охолони чуток, шибко не кипятись, шибко-то. Извиняй, говорит, погорячился, может, и не отец. Саша Соколов. Между собакой и волком

2.1. (что-л.) создав что-л. новое, сделать это самостоятельным и дееспособным, *что осмысляется как воспитание детей* ❖ укрепить что-л., развить что-л. 🗐 Идея была такая: организовать школу, *поставить* ее *на ноги* и через два года передать проект директору-москвичу. Два года еще не истекло, но уже сейчас ясно, что в Англию уезжать рано. Потому что если не перевести проект на самофинансирование, он зачахнет и погибнет в паутине нашей равнодушной инертности. А ведь цель — показать модель и на ее основе организовать сеть подобных школ по всей Москве. Московский комсомолец

3. (кого-л./что-л.) заставить кого-л. (человека, людей или организацию) активно, энергично действовать, *что описывается как придание человеку устойчивого вертикального положения, рассматриваемого как норма и являющегося необходимым условием для начала движения и активной деятельности* ❖ активизировать кого-л./что-л. 🗐 Сам я *поднял на ноги* всю временно находившуюся на свободе привокзальную шоблу жулья, проституток, «уносильщиков» чужих чемоданов, разгонщиков, барыг, официанток и так далее. Юз Алешковский. Призрак в белом халате.

4. (кого-л.) заставить проснуться кого-л., внезапно прервав его сон, *что осмысляется как придание человеку устойчивого вертикального положения, рассматриваемого как норма и являющегося необходимым условием для начала движения и активной деятельности* ❖ разбудить кого-л. 🗐 Надо сказать, что проводы прошли хорошо, хотя несколько затянулись. Последнего гостя мы вытолкали без четверти три ночи, а четверть седьмого утра жена уже *подняла* меня *на ноги*. В. Войнович. Москва 2042.

Первое, второе, третье и четвертое значения представляют собой фрагменты радиальной полисемии, которые мотивированы непосредственно внутренней формой идиомы. При этом в первом и втором значении профилируется один признак образа ‘устойчивое вертикальное положение человека’, противопоставленное горизонтальному, а в третьем и четвертом значениях другой — ‘устойчивое вертикальное положение человека являющееся необходи-

[4] Отметим, что в Словаре нумерация значений идиом со смешанной полисемией не отражает их иерархию.

мым условием для начала движения'. Значения **1.1** и **2.1** производны от, соответственно, первого и второго значений и представляют собой два цепочечных элемента в структуре идиомы.

VI. Таким образом, рассмотренный материал показывает, что структура многозначных идиом может иметь цепочечный, радиальный и смешанный вид. В этом идиоматика сходна с лексикой. Отличие же заключается в том, что во фразеологии в силу мотивированности значения идиомы ее внутренней формой увеличивается количество единиц, охватываемых топологической классификацией, т. к. становится возможным установить тип многозначности (радиальной или цепочечной) для идиом с двумя значениями. Не совпадает и распространенность типов многозначности: для лексики это смешанный тип, для фразеологии — радиальный. Дальнейшее изучение различных аспектов полисемии идиом позволит установить как специфические особенности идиоматики, так и признаки, сближающие фразеологические и лексические единицы.

Литература

1. *Апресян Ю. Д.* Лексическая семантика (синонимические средства языка) // М.: 1974.
2. *Апресян Ю. Д.* Исследования по семантике и лексикографии. Том I: Парадигматика // М.:2009.
3. *Баранов А. Н., Вознесенская М. М., Добровольский Д. О., Киселева К. Л., Козеренко А. Д.* Фразеологический объяснительный словарь русского языка // М.: 2009.
4. *Баранов А. Н., Добровольский Д. О.* Аспекты теории фразеологии // М.: 2008.
5. *Зализняк Анна А.* Многозначность в языке и способы ее представления // М.: 2006.
6. *Кустова Г. И.* Типы производных значений и механизмы языкового расширения // М.: 2004.
7. *Падучева Е. В.* Динамические модели в семантике лексики // М.: 2004.
8. *Розина Р. И.* Семантическое развитие слова в русском литературном языке и современном сленге // М.: 2005.
9. *Телия В. Н.* Русская фразеология. Семантический, прагматический и лингвокультурологический аспекты // М.: 1996.

О распознавании жестов языка глухих[1]

About recognition of sign language gestures

Воскресенский А. Л. (avosj@yandex.ru), независимый исследователь,
Ильин С. Н. (mail@mocaprus.ru), «Академия фантазий», Москва
Milos Zelezny (zelezny@kky.zcu.cz), University of West Bohemia, Plzeň, Czech Republic

Обсуждаются задачи восприятия и распознавания жестов русского языка глухих в системе автоматизированного сурдоперевода. Предлагаются новый подход к морфологии жестов, метод выделения отдельных жестов жестового высказывания. Предлагается рабочее определение понятия «понимание текста».

Введение

Создаваемая система автоматизированного сурдоперевода [1] должна быть способна не только переводить воспринятый текст (и речевое сообщение) в жестовые высказывания, но и преобразовывать жестовые высказывания в слова и фразы. С последней задачей, как показывает практика, во многих случаях не справляются люди-сурдопереводчики [2]. Одной из причин этого может быть ошибочное разделение жестового выражения на составляющие жесты.

Турецкие участники проекта «Информационный киоск» [3], пытаются найти границы жестов, сопоставляя их с параллельно произносимыми диктором-сурдопереводчиком словами. Но этот подход ограниченно применим лишь при обработке «кальки» [4], он бесполезен при восприятии жестовой речи, в которой поток жестов не совпадает с потоком слов.

При создании систем машинного сурдоперевода возникает дополнительная проблема восприятия жеста системой. Например, в работах французских исследователей [5] возникали проблемы распознавания жеста при съемке одной камерой.

В данной работе описываются подходы к распознаванию жестов, лежащие в основе создания составной части системы сурдоперевода: подсистемы перевода «жесты — текст». В разделе 1 дается краткое описание используемого способа съемки жестов и преобразования их в 3D-анимацию, предлагаемый метод распознавания жестов; в разделе 2 приводятся соображения о морфологии жеста, основанные на исследовании жестов, представленных в словаре RuSLED [6], и методе разделения жестового высказывания на отдельные жесты; раздел 3 относится к пониманию словесных и жестовых высказываний.

1. Средства отображения, восприятия и распознавания жестов

Жест языка глухих представляет собой комбинацию конфигураций пальцев рук (одной или двух), положений рук относительно тела говорящего (с учетом направления движения рук) и сопутствующей мимики, передающей эмоциональную составляющую. При общении на «кальке» мимика включает в себя и артикуляцию, соответствующую произнесению (зачастую беззвучно) соответствующего слова. В «истинной» жестовой речи артикуляция используется для указания значения омонимичного жеста.

Передача столь сложной комбинации, осуществляемой в пространстве и времени, весьма затруднительна для записи. Для записи жестов используются различные варианты нотации, например Гамбургская система нотации (HamNoSys, http://www.sign-lang.uni-hamburg.de/projects/hamnosys.html). В России используется нотация, предложенная Л. С. Димскис [7].

Нотация HamNoSys использовалась в европейском проекте eSign (http://www.sign-lang.uni-hamburg.de/esign/) для управления движениями

[1] Данное исследование проводится при поддержке Министерства Образования, Молодёжи и Спорта Чешской Республики в рамках совместного проекта DIMAS-CZ, No. ME08106.

аватара Guido, демонстрирующем жесты. Мимику этот аватар не отображал.

В настоящее время эта нотация используется в проекте «Информационный киоск», в котором от России участвует СПИИРАН. Аватар, демонстрирующий жесты, разработан в университете Западной Богемии (Чешская республика). Несмотря на то, что дополненная версия HamNoSys имеет знаки для отображения некоторых элементов мимики, при управлении аватаром они не используются. Версия аватара, модифицированная СПИИРАН, артикулирует русские слова.

Словарь русского жестового языка RuSLED дополнен функцией поиска жеста по его описанию. Задача состоит в том, что нужно найти жест, который человек видел, но не знает его значение. Для этого используется упрощенная нотация, зашитая внутри словаря, для пользователя доступны списки возможных значений — текстовые для описания места исполнения жеста, текст с рисунком — для конфигураций пальцев. На основе выбранных пользователем значений формируется поисковый запрос и выдается набор жестов, отвечающих этому запросу, из которого пользователь выбирает интересующий его жест.

Демонстрация жестов в новой версии словаря осуществляется анимированным персонажем — аватаром, для записи жестов используется методика «захвата движений» (motion capture). Запись жестов проводится студией «Академия фантазий» (www.mocaprus.ru). Движения демонстратора, фиксируемые с помощью 12 камер и отражателей на костюме (рис. 1), преобразуются в 3D-модель (рис. 2), используемую для формирования облика аватара, который может быть помещен в любую сцену (рис. 3).

Движения пальцев демонстратора фиксируются с помощью специальных перчаток. Для снятия мимики и артикуляции используются фиксируемые на лице отражатели (рис. 4). Их сигналы преобразуются в трехмерную модель мимики лица (рис. 5).

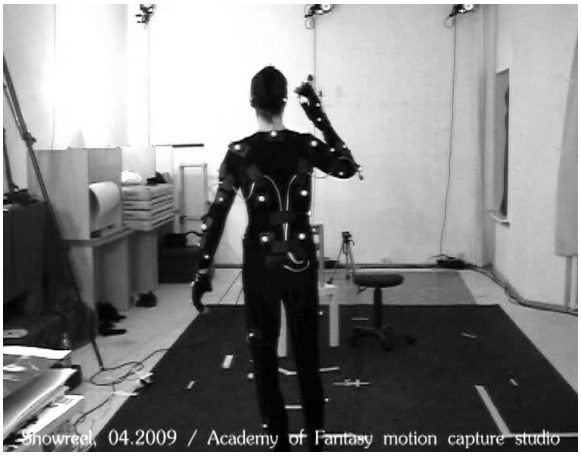


Рис. 1. Демонстратор в костюме с отражателями

Рис. 2. 3D-модель

Рис. 3. Аватар, построенный на основе 3D-модели и помещенный в виртуальную сцену

Реализация подобного преобразования позволит существенно ускорить наполнение словаря за счет использования нескольких сурдопереводчиков для демонстрации жестов, сохраняя при этом единство действия, выраженное единым обликом виртуального демонстратора жестов. Сформированный таким образом словарь позволит компоновать жестовые высказывания из хранящейся в словаре коллекции жестов, сохраняя, как указывалось выше, единство действия, что важно для восприятия жестовых высказываний человеком.

Студийная запись жестов, позволяя формировать исходный словарь, очевидно, не может быть средством коммуникации с глухими людьми.

Для распознавания жестов, например, при преобразовании жестовых высказываний в текст, предполагается разработка средств преобразования снятых видеокамерой растровых изображений сурдопереводчика в векторное изображение. Это преобразование включает в себя распознавание существенных для данной задачи деталей изображения: голова, кисти рук (и положение каждого пальца),

туловище. Указанные детали изображения преобразуются в эллипсы и прямоугольники, координаты которых сопоставляются с параметрами скелета виртуального демонстратора (аватара).

Рис. 4. Демонстратор с наклеенными на лицо отражателями

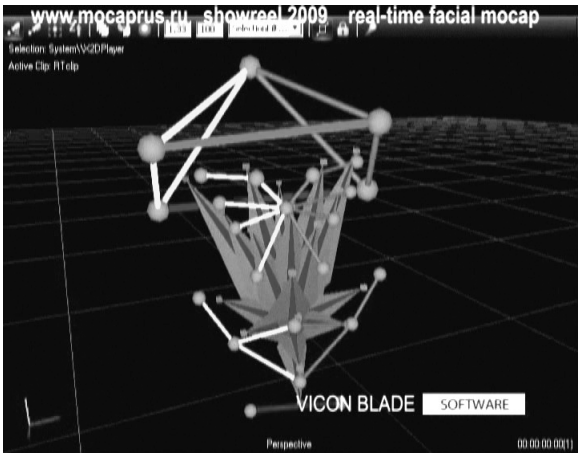


Рис. 5. Трехмерная модель лицевой мимики

Преобразование воспринятых изображений в векторную форму позволяет существенно сократить требования к объему памяти интеллектуальной системы и ускорить процедуры сравнения с эталонами. Представляется, что подобная методика может быть использована не только для распознавания жестов, но и для более широкого круга задач.

Методы преобразования изображения, которые предполагается использовать, близки к тем, которые используются при предварительной обработке изображений в системах распознавания символов (см., например, [8]).

Наиболее близким аналогом являются работы отдела Artemis Национального института телекоммуникаций Франции (http://www-artemis.it-sudparis.eu/web2.0/index.php?pid=1).

Отличием является то, что в разработках отдела Artemis ставится задача наиболее близкого отображения облика сканируемых объектов, тогда как в данном проекте внешний вид сканируемого объекта заменяется заранее заданным обликом аватара, но при этом ставится задача наиболее точного отображения положений пальцев, рук, туловища в целом. Информацию о точном положении в пространстве, например, рук, отсутствующую в двумерном растровом изображении, полученном от одной видеокамеры, планируется извлекать из знаний о возможных и допустимых взаимных положений различных частей тела человека. Для точного определения позы аватара будут применяться соответствующие геометрические построения, обеспечивающие наиболее близкое совпадение проекций аватара на плоскость отображения с исходным растровым изображением.

В случае достаточно надежного распознавания жестов с помощью одной видеокамеры (желательно добиться качественного распознавания с помощью типовых веб-камер) и создания системы сурдоперевода возможно будет обеспечить оперативную коммуникацию глухих с представителями администрации и общественности, что является одной из функций «электронного правительства».

2. О морфологии жестов

Считается, что впервые морфологию жестового языка описал W. Stockoe [9]. Соответственно, всякий жест этого языка (функционально близкий морфеме) складывается из хирем (от греч. χείρ — рука), делящихся на три класса — табы указывают на место исполнения жеста, дезы — на конфигурации руки, а сиги — на характер движения. Хиремы функционально эквивалентны фонемам, но в отличие от фонем, выстраивающихся в морфеме в линейную последовательность, в жесте-морфеме одновременно присутствует хирема каждого из трех классов. Общее количество хирем сопоставимо с числом фонем в звуковых языках — в ASL (американском жестовом языке) имеется 12 табов, 19 дезов и 24 сига, в шведском жестовом языке, соответственно, 18, 22 и 24, в языке глухих южной Франции — 16, 17 и 20 и т. д. W. Stockoe разработал для ASL систему записи жестов как последовательности таба, деза и сига. Эти положения лежат в основе разработки систем нотационной записи жестов.

Однако, как значения слов не всегда определяются составляющими морфами, завися от контекста, так и при анализе жестовых высказываний нужно учитывать, что многие жесты являются составными, содержат в себе комбинацию нескольких жестов и предшествующих дактильных знаков, модифицирующих значение данного жеста. Несколько примеров приведено в табл. 1.

Таблица 1. Примеры составных жестов

Наименование жеста	Предшествующие дактилемы	Составляющие жесты
абитуриент	—	учиться, войти
абрикос	а	Имитация разламывания плода
автомат (робот)	—	механизм, задание
азбука	а, б, в	—
акварель	а	краска
алый	—	красный, яркий
алюминий	а	металл
античность	а	старый, было
аттестат	а	печать, выдать
аудитория (помещение)	—	учиться, комната
аудитория (слушатели)	—	слушать, все
бронза	б	металл
девочка	—	женщина, маленький
душа	д	я
ненаглядный	—	смотреть, любить
озеро	—	вода, площадь (место)
океан	о	море
она (об отсутствующей при разговоре)	—	женщина, вы
пословица	—	говорить, предложение
раскаиваться	—	ошибка, ум
снег	—	белый, падать

Когда нужно передать, например, падежные окончания, вслед за жестом показываются соответствующие дактилемы.

Число составных жестов велико, но статистику приводить в настоящее время преждевременно. Это связано со сравнительно малым объемом имеющихся словарей жестов, что может привести к существенному смещению статистических оценок.

Приведенный весьма ограниченный (и, возможно, не самый показательный) набор примеров показывает, что жестовый язык изобилует образными и идиоматическими выражениями, значение которых не всегда совпадает с суммой значений составных элементов. Возможно, именно это приводит к ошибкам сурдопереводчиков при переводе жестов в текст и речь.

Жестовая речь не содержит пауз между отдельными жестами. Паузами разделяются лишь фразы. Это привносит дополнительные сложности при осуществлении автоматизированного сурдоперевода, напоминающие те, которые встречаются при разработке систем распознавания слитной речи.

Учитывая составной характер жестов, разделение жестовой фразы на отдельные жесты следует вести, выбирая из словаря соответствующие жесты, имеющие наибольшую длину, и анализируя семантику получаемого высказывания. В случае, если его значение не соответствует дискурсу, можно приступить к поочередному расщеплению «длинных» жестов на составные элементы, пытаясь получить высказывание, содержание которого соответствует дискурсу.

Это лишь предварительные предположения, которые подлежат экспериментальной проверке.

3. Сопоставление значений словесных и жестовых высказываний

При сурдопереводе, как и в других случаях перевода с одного языка на другой, основной задачей является правильная передача значения переводимого сообщения. Для этого необходимо понимать исходное сообщение, что является сложной психологической задачей [10].

В данной работе принято следующее рабочее определение понимания текста:

Результатом понимания текста должно быть опознание объектов, описанных в тексте, их пространственного положения, а также изменений их характеристик, действий и положения в соответствии с изменением времени текста.

Это определение было выработано для задачи перевода текста в жесты [6]. Однако оно действительно и для задачи перевода жестовых высказываний в текст. Даже еще в большей степени, учитывая необходимость определения объектов, описываемых одинаковыми жестами. Так, например, местоимения «*он*», «*она*», «*оно*» передаются одним и тем же жестом (в случае присутствия в месте разговора). Для нахождения правильного значения в этом случае необходим анализ всего дискурса [11].

Опознание объектов включает в себя не только выделение именных групп, описывающих тот или иной объект, но и распознавание того, встречался ли ранее в тексте данный объект, совпадают ли объекты, имеющие одинаковые имена [12].

В [13] указывается, что классическая линейная схема процесса обработки речи (рис. 6) с 60-х годов двадцатого столетия признается психолингвистами нереалистичной. Её предлагается заменить схемой, учитывающей взаимодействие различных этапов обработки и обратные связи между ними (рис. 7).

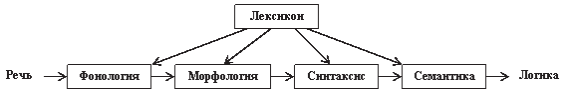


Рис. 6. Классическая схема обработки речи

При переводе с жестового языка из полученного в результате распознавания набора наименований жестов необходимо генерировать грамматически правильные фразы, отражающие содержание жестового высказывания. Это многоступенчатый процесс анализа получаемых вариантов, в ходе которого возможны возвращения на предшествующие этапы обработки, вплоть до выбора иного разделения жестовой фразы на составляющие фразы. Схема подобного процесса может быть близка к схеме, изображенной на рис. 7.

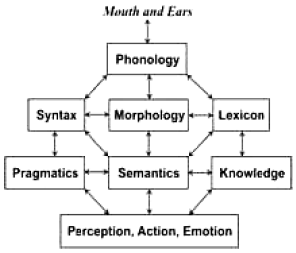


Рис. 7. Когнитивистский подход к пониманию речи

Как показано в [14], разрешение дейктической референции при обработке жестовых высказываний требует знания о пространственном положении объекта. Анализ пространственного положения объектов, описываемых в тексте, подразумевает способность интеллектуальной системы представлять визуальные образы, сопоставлять их друг с другом, расставлять на пространственной шкале «ближе» — «дальше». В [15] описан подход к созданию прототипа интеллектуальной системы, способной распознавать и сопоставлять визуальные и текстовые объекты. При этом каждая подсистема (в том числе концептуальной графики, обработки и синтеза текста) участвует в работе совместно с другими подсистемами, представляя общий результат на едином концептуальном языке прикладной онтологии, являющейся общим описанием окружающего мира для всех подсистем.

Заключение

Описываемая работа далека от завершения, однако и на данном этапе ожидаются полезные результаты. Использование методов motion capture позволяет быстро формировать словарь жестового языка. Так, например, во время одной из пробных съемок за час работы были сняты три серии по 30 отдельных жестов. Это позволяет надеяться, что пополнение словаря жестов может вестись со скоростью до 50 жестов в день. В качестве демонстраторов жестов используются глухие носители жестового языка, что позволяет снять замечания по содержанию словаря RuSLED, вызванные тем, что включенные в него жесты демонстрировались слышащими людьми, не вполне правильно изображавшими некоторые жесты.

Формирование объемного словаря жестового языка позволит уточнить сведения по морфологии жестов, а также послужит основой формирования библиотеки шаблонов для планируемой работы по распознаванию жестов.

Благодарность

Авторы выражает искреннюю благодарность Н. А. Чаушьян, чьи замечания позволили обратить внимание на содержащиеся в словаре RuSLED ошибки и принять меры к их исправлению. Её весьма полезные замечания по жестовому языку позволили обратить внимание на особенности, не всегда заметные исследователю, не являющемуся носителем языка.

Литература

1. *Voskressenski A.* Signs and speech: two forms of human communication // Proceedings of the Ninth International Conference «Speech and Computer» SPECOM'2004. Saint-Petersburg, Russia, 2004, P. 666–669.
2. *Овсянникова Л. А.* Проблемы жестового перевода на телевидении // Русский жестовый язык и проблемы перевода: Материалы конференции. — М., 2001.
3. *Hruz M., Campr P., Karpov A., Santemiz P., Aran O. and Zelezny M.* Input and Output Modalities Used in a Sign-Language-Enabled Information Kiosk // Proceedings of the 13-th International Conference "Speech and Computer" SPECOM'2009. — St. Petersburg: SUAI, 2009. — P. 113 — 116.
4. *Зайцева Г. Л.* Жестовая речь. Дактилология: Учебное пособие для ВУЗов. — М.: ВЛАДОС, 2000.
5. *T. Zaharia, F. Prêteux.* Video archiving and sign language indexation within the AMIS platform // Proceedings IASTED Conference on Signal Processing, Pattern Recognition and Applications (SPPRA'02), Crete, Greece, 2002, p. 396–401.
6. *Воскресенский А. Л., Гуленко И. Е., Хахалин Г. К.* Словарь RuSLED как инструмент семантических исследований. // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2009» (Бекасово, 27–31 мая 2009 г.). Вып. 8 (15).– М.: РГГУ, 2009. — С. 64–68.
7. *Димскис Л. С.* Изучаем жестовый язык: Учеб. пособие для студ. дефектол. фак. высш. пед. учеб. заведений. — М.: Издательский центр «Академия», 2002. — 128 с.
8. *Масалович А. А.* Кластеризация изображений графем на основе непрерывного граничноскелетного представления. // Математические методы распознавания образов ММРО-12: Доклады 12-й Всероссийской конференции. — М., 2005. — С. 374–378.
9. *Stockoe W. C.* Sign language sructure: An outline of the usual communication system of the American Deaf / Buffalo 14, New York University of Buffalo, 1960. — 79 p.
10. *Лурия А. Р.* Понимание компонентов речевого высказывания // Языки сознание / Ред. Е. Д. Хомской. — М.: Изд. МГУ, 1979. — С. 217–226.
11. *Ван Дейк Т. А., Кинч В.* Стратегии понимания связного текста. // Новое в зарубежной лингвистике. — Вып. 23. Когнитивные аспекты языка. — М., 1988.
12. *Kazi Z., and Ravin. Y.* Who's Who? Identifying Concepts and Entities across Multiple Documents.// Proceedings of the 33rd Hawaii International Conference on System Sciences — 2000. (0-7695-0493-0/00).
13. *Majumdar A., Sowa J., and Stewart. J.* Pursuing the Goal of Language Understanding // Proceedings of the 16th ICCS /P. Eklund and O. Hammerlé, eds. — LNAI 5113, Springer, Berlin, 2008, pp. 21–42.
14. *Кибрик А. А., Прозорова Е. В.* Референциальный выбор в русском жестовом языке. // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2007» (Бекасово, 30 мая — 3 июня 2007 г.) / Под ред. Л. Л. Иомдина, Н. И. Лауфер, А. С. Нариньяни, В. П. Селегея. — М.: Изд-во РГГУ, 2007. — С. 220–230.
15. *Хахалин Г. К., Воскресенский А. Л.* Мультизадачное использование прикладной онтологии. // Одиннадцатая национальная конференция по искусственному интеллекту с международным участием КИИ-2008 (28 сентября — 3 октября 2008 г., г. Дубна, Россия): Труды конференции. В 3 т. — М.: ЛЕНАНД, 2008. — Т. 1. С. 112–123.

Автоматический поиск и классификация однословных терминов в корпусе предметной области с использованием логарифмической меры сходства с неспециализированным корпусом

Automatic detection and classification of single-word terms in a specific domain corpus using log-likelihood similarity with general purpose corpus*

Гельбух А. Ф. (www.gelbukh.com),
Сидоров Г. О. (www.cic.ipn.m/~sidorov),
Лавин-Вийа Э.
Лаборатория естественного языка и обработки текста,
Центр Компьютерных Исследований (CIC),
Национальный Политехнический Институт (IPN), г. Мехико, Мексика

Чанона-Эрнандес Л.
Инженерный факультет (ESIME),
Национальный Политехнический Институт (IPN), г. Мехико, Мексика

В статье представлен метод поиска однословных терминов в корпусе предметной области, использующий логарифмическую меру сходства (log-likelihood) с неспециализированным корпусом. Также проводится автоматическая классификация полученных терминов на основе меры сходства по косинусу угла.

1. Введение

Автоматическое построение онтологий для специфических предметных областей, или, по крайней мере, построение списка терминов и гипотез о возможных отношениях между ними для последующей ручной обработки, является важной и актуальной задачей современной компьютерной лингвистики. Это связано с трудностями ручной разработки онтологий: необходимость в экспертах для каждой предметной области, невысокая скорость их работы, большие затраты, а также субъективность полученных данных (см. например, Uschold & Grunninger, 1996). Эта область является достаточно активно развивающейся. В настоящее время не существует какого-либо общепринятого метода построения онтологий.

В качестве первого шага такого построения является полезным извлечение однословных терминов из текстов или слов, которые являются частью многословных терминов. В дальнейшем из этих однословных терминов могут быть сформированы многословные термины. Заметим, что большинство терминов состоит из нескольких слов, и большинство существующих методов сразу пытается их извлекать, измеряя степень совместной встречаемости составляющих их слов. Казалось бы, может быть проще набросать проект выделения таких терминов вручную вместо разработки программы. Мы уже упоминали субъективность такого подхода. Кроме того, для одной-двух предметных областей это может быть проще, то если их уже сто или тысяча, то наверное проще применять программу автоматической обработки.

* Work done under partial support of Mexican Government (CONACYT, SNI) and National Polytechnic Institute, Mexico (SIP, COFAA, PIFI; projects SIP 20091587, 20090772, 20100773, 20100668).

В данной статье мы представляем метод, являющийся модификацией метода описанного для китайского языка в (Tingting He *et al*., 2006), и эксперименты построения проекта онтологии на материале испанского языка. Мы работали с предметной областью «Информатика». Метод является достаточно универсальным и может быть применен к широкому классу языков (в том числе, к русскому) с соответствующими изменениями в предобработке. Метод дает достаточно хорошие предварительные результаты, выделяя однословные термины предметной области и объединяя их в классы.

Далее в статье мы сначала описываем предлагаемый метод (предобработка, поиск терминов, мера их сходства и классификация). Затем приводятся данные о проведенном эксперименте, и в заключение делаются выводы.

2. Предлагаемый метод

Входными данными метода является корпус специфической предметной области (мы работали с текстами по информатике). Также метод предполагает наличие неспециализированного корпуса для его использования при сравнении. Заметим, что размер этих двух корпусов может быть очень разным.

Предлагаемый метод состоит из четырех основных шагов: предобработка и подготовка данных, поиск терминов, вычисление меры их сходства и объединение терминов в классы.

Метод использует традиционную в информационном поиске векторную модель представления набора документов, когда документы и слова из них являются двумя измерениями матрицы, которая содержит частоты данного слова в данном документе. Заметим, что эквивалентным является представление в виде таблицы содержащей набор: *слово-документ-частота*, например,

Табл. 1. Представление данных в виде таблицы

Слово	Документ	Частота
software	14	3
software	16	3
software	20	12
и т. п.	…	…

Очевидно, что в этом случае нет необходимости добавлять записи для документов, в которых данного слова просто нет, т. е. в которых слово имеет частоту 0. Таких записей просто не будет в этой таблице.

Метод был модифицирован в следующих аспектах по сравнению с описанным в (Tingting He *et al*., 2006) методом для китайского языка: он применяется к индоевропейскому языку (испанскому) с соответствующим изменением предобработки; мы не используем обогащение словаря дополнительными ресурсами (типа WordNet), что является существенной частью исходного метода, при этом метод продолжает оставаться достаточно надежным; мы изменили формулу вычисления разницы между корпусами — вместо вычисления разницы по алгоритму *loglikelihood test*, мы вычисляем разницу по другому алгоритму, тоже основанному на *loglikelihood*, который вычисляет «расстояние» между элементами (Rayson *et al*., 2004; Dunning, 1993). Заметим, что этот алгоритм как раз и предназначен для вычисления сходства корпусов.

2.1. Предобработка и подготовка данных

На этапе предобработки в документах выделяются слова. В нашем случае мы игнорировали знаки препинания, специальные символы, числа. Все слова приводятся в один регистр.

Все слова лемматизируются. Мы пользовались лемматизатором для испанского языка, разработанным в нашей лаборатории. Для русского языка также существуют доступные лемматизаторы (см. например, Gelbukh and Sidorov, 2005).

Кроме того, мы отфильтровываем все служебные слова (предлоги, союзы, вспомогательные глаголы, и пр.), так как заранее известно, что они не являются терминами.

Для полученных лемм подсчитываются их частоты в каждом документе и заносятся в матрицу.

Эта процедура проделывается отдельно для каждого корпуса. В результате мы получаем два матрицы (или таблицы), которые представляют весь корпус исходных текстов и весь неспециализированный корпус.

2.2. Поиск терминов с использованием логарифмической меры сходства

Как уже было сказано, метод использует два корпуса: корпус специфической предметной области и неспециализированный корпус. Основная идея состоит в сравнении взвешенных частот слов в двух корпусах, и если какое-либо слово гораздо чаще присутствует в корпусе предметной области, то это вероятный термин.

Заметим, что в (Tingting He *et al*., 2006) указано, что логарифмическая мера сходства дает лучшие результаты чем гораздо более традиционная мера TF/IDF.

Мы также применили логарифмическую меру сходства, но не в варианте теста (*loglikelihood test*, см. www.wikipedia.org), а в варианте, который предназначен для сравнения корпусов (Rayson *et al*., 2004).

Для каждого слова, вычисление проводилось по следующей формуле:

$$G = 2 * \left(\left(fr_{domain} * \log\left(\frac{fr_{domain}}{frExpected_{domain}} \right) \right) + \left(fr_{general} * \log\left(\frac{fr_{general}}{frExpected_{general}} \right) \right) \right)$$

где:

$frExpected_{domain}$ и $frExpected_{general}$ ожидаемые частоты в корпусе предметной области и в неспециализированном корпусе соответственно;

fr_{domain} и $fr_{general}$. реально наблюдаемые частоты в корпусе предметной области и в неспециализированном корпусе соответственно.

Табл. 2. Пример вычислений веса терминов

Слово	fr_{domain}	$fr_{general}$	$frExpected_{domain}$	$frExpected_{general}$	G
socket	1	0	0,010286744	0,989713252	9,153798
sofisticado (сложный)	5	169	1,789893508	172,210113500	3,912798
soft	1	12	0,13372767	12,866271970	2,351035
software	430	831	12,97158432	1248,028442000	2334,961
*sofware***	2	2	0,041146975	3,958853006	12,8037
sol (солнце)	2	933	9,618105888	925,381897000	−9,016687
solamente (только)	20	1714	17,83721352	1716,162842000	0,254846

** Это слово написано с ошибкой в корпусе. Правильно *software*

Для вычисления ожидаемых частот используются следующие формулы.

Обозначим через *R_fr* такое отношение частот:

$$R\_fr = \frac{fr_{domain} + fr_{general}}{size_{domain} + size_{general}}$$

где $size_{domain}$ и $size_{general}$ размеры соответствующих корпусов, вычисленные в количестве слов. Тогда

$$frExpected_{domain} = size_{domain} * R\_fr$$

$$frExpected_{general} = size_{general} * R\_fr$$

Следующий важный шаг в нашем алгоритме поиска терминов состоит в следующем. Заметим, что значения полученные по формуле, не различают, к какому корпусу относится предполагаемый термин (т. е. формула симметрична относительна корпусов). Нас это не устраивает, потому что мы ищем термины в предметной области, а не в неспециализированном корпусе. Для принятия во внимание этого явления, вычислим дополнительно относительные частоты слов в каждом корпусе и будем рассматривать только те слова, у которых больше относительная частота в корпусе предметной области. Для этого, если относительная частота в корпусе предметной области меньше, чем в в неспециализированном корпусе, например, умножим результат, полученный по вышеуказанной формуле, на −1.

В Таблице 2 приведены примеры полученных вычислений.

Например, у слова *software* (*программное обеспечение*) получился очень высокий вес. Вес же слова *sol* (*солнце*), хотя и достаточно высок, но был умножен на −1, так как его относительная частота больше в неспециализированном корпусе.

По окончании этого этапа метода, у нас есть список слов с их весом в корпусе предметной области, который соответствует их вероятности быть терминами этой области. Остается открытым вопрос о том, по какому порогу провести границу между терминами и не-терминами (см. раздел Эксперименты).

2.3. Вычисление меры сходства терминов по косинусу угла

Следующий этап метода состоит в вычисление меры сходства терминов по косинусу угла (см. например, cosine similarity в wikipedia). Эта мера будет использоваться для классификации терминов на следующем шаге. В данном вычислении используется стандартная формула из информационного поиска. В качестве данных мы, как это обычно делается, используем частоты отобранных слов. Обычно эта мера сходства отражает совместную встречаемость слов в одном документе, нам кажется, что эта интерпретация применима и в нашем случае.

Естественно, вычисления проводятся только для слов, отобранных на предыдущем этапе алгоритма.

$$\cos(x,y) = \frac{A \cdot B}{|A||B|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}$$

где:

n это количество документов в корпусе предметной области;

x_i и y_i это частота слов x и y в документе i.

В Таблице 3 приведены примеры вычислений сходства разных слов со словом *software (программное обеспечение)* в алфавитном порядке.

Табл. 3. Пример вычислений сходства терминов для слова ***software***

Слово	Сходство
algoritmo (алгоритм)	0,032
almacén (склад)	0,018
almacenamiento (хранение)	0,044
almacenar (хранить)	0,086
aplicación (приложение)	0,420
archivo (файл)	0,203
arpanet (arpanet)	0,031
arquitectura (архитектура)	0,271
artificial (искусственный)	0,029
base (основа)	0,119
bioinformática (биоинформатика)	0,073
cálculo (вычисление)	0,055
característica (характеристика)	0,220
ciencia (наука)	0,071
circuito (плата)	0,018
código (код)	0,759
compilador (компилятор)	0,185
componente (компонента)	0,171
и т. д.	...

Можно заметить, что наибольшее сходство у слова *software* среди слов, отобранных как потенциальные термины в области информатики, со словом *код* (0,759) и наименьшее со словом *биоинформатика* (0,073).

2.4. Классификация терминов

В качестве алгоритма классификации мы использовали стандартный алгоритм *k-means* (www.wikipedia.org).

Этот алгоритм получает в качестве входного параметра количество классов k, которые должны быть сформированы. Алгоритм состоит в следующем:

- Случайным образом выбираются k терминов в качестве центров классов.
- Каждый оставшийся термин классифицируется на основе сходства с центром класса (по наибольшему сходству, см. выше).
- Заново высчитываются центры классов.
- Два предыдущих шага повторяются до тех пор, пока есть какие-то изменения в результатах.

В результате получаются k классов, состоящих из наболее похожих терминов. См. пример в Таблице 4.

Табл. 4. Пример результатов классификации в одном из экспериментов

Центр	Элементы класса	Сходство с центром
describir (описать)	*describir (описать)*	1,000
	solucionar (решить),	0,729
	arquitectura (архитектура компьютера),	0,530
	matemática (математика),	0,509
	interfaz (интерфейс),	0,449
	patrón (образ, шаблон),	0,439
	diseñar (разрабатывать),	0,427
	diseñador (разработчик)	0,335
disco (диск)	*disco (диск),*	1,000
	dvd (DVD),	0,947
	disquete (дискета),	0,934
	rom (ROM),	0,933
	cd (CD),	0,928
	usb (USB),	0,899
	flash (флеш),	0,877
	almacenamiento (хранение),	0,872
	óptico (оптический),	0,840
	velocidad (скорость),	0,753
	soportar (поддерживать)	0,591
и т. п.	...	...

3. Эксперименты

3.1. Данные и параметры

В наших экспериментах мы использовали следующие данные (на испанском языке). Для сравнения в качестве неспециализированного корпуса были выбраны выпуски газеты Excélsior (Мексика) конца 90-х годов, всего 1 365 991 слов.

В качестве корпуса предметной области были взяты страницы из *wikipedia*, связанные с информатикой: *информатика, программное обеспечение (software), программирование*, и т. п. Всего было загружено 26 страниц, содержащих 44 495 слов. В принципе, для экспериментов можно взять любую коллекцию текстов и это не требует какого-либо дополнительного обоснования.

После нескольких предварительных экспериментов, мы выбрали порог для алгоритма поиска терминов в 270 терминов. В статье (Tingting He *et al.*, 2006) был выбран порог в 216 терминов, при том, что дополнительно производился анализ отношений в ки-

тайском аналоге WordNet. В качестве порога на количества классов, мы выбрали порог в 19 классов, для сравнения в указанной статье был выбран порог в 20 классов. Эти параметры легко варьировать в экспериментах. Порог был выбран эмпирически, т. е., мы пробовали разные значения и остановились на пороге, который дает интуитивно лучшую классификацию.

3.2. Результаты

После применения нашего метода к указанным корпусами с данными параметрами, были получены следующие результаты. Приведем несколько терминов, которые получили наибольший вес в соответствии с алгоритмом поиска терминов, см. Таблицу 5.

Табл. 5. Термины области «Информатика» с наибольшим весом

Термин	Log-likelihood
Dato (данные)	1506
Computador (компьютер)	863
Circuito (плата)	467
Memoria (память)	384
Señal (сигнал)	372
Secuencia (последовательность)	353
Computación (вычисление)	351
Información (информация)	346
Dispositivo (устройство)	342
Algoritmo (алгоритм)	341
Electrónico (электронный)	322
Base (основа)	319
Diseñar (разрабатывать)	307
Utilizar (использовать)	282
и т. д.	…

Интересный вопрос состоит в том, нужно ли включать в этот список глаголы, т. к. большинство глаголов в этом случае являются просто лексическими функциями к соответствующим существительным. В случае если было бы принято решение об исключении глаголов, это легко сделать, т. к. был проведен морфологический анализ и лемматизация.

После применения алгоритма классификации к полученным терминам с порогом на 19 классов, были получены следующие результаты, см. Таблицу 6. Для простоты не будем приводить данные о сходстве слов с центром класса. Первое слово в каждом классе является его центром. Как можно заметить, некоторые слова приводятся по-английски (*for, to, DAQ*, и т. д.), как они были представлены в исходных текстах.

В таблице мы зачеркнули слова, которые явно НЕ являются терминами с нашей точки зрения и подчеркнули глаголы, которые мы не будем использовать для подсчетов, т. к. необходимость их включения в онтологию небесспорна. Оставляем открытым вопрос о том, нужно ли исключить также и отглагольные существительные.

Среди отобранных слов присутсвует большое количество слов общенаучной лексики, типа *анализ, система, модель, наука, теория,* и пр. Строго говоря, они не должны присутствовать в онтологии выбранной предметной области, но с другой стороны они все-таки являются научными терминами. Вероятно, их можно отфильтровать проделав аналогичное описанному сравнение корпусов, но уже из двух различных предметных областей. Мы выделили эти слова в Табл. 6 курсивом. Как обычно, четкого критерия выделения таких слов нет, поэтому мы руководствовались следующим принципом: «если у слова есть в данной области какое-то специфическое значение, отличное от общенаучного, то он считается термином», например, *объект* может быть термином в программировании (*объектно-ориентированное*; по-испански, *ориентированное на объекты*). Или слово *протокол*, когда существуют *протоколы передачи данных*, и пр.

Некоторые слова оказались в двух классах, потому что их расстояние до двух центров было одинаковым. Кроме того, необходимо пояснить, что когда система морфологического анализа не имела в в словаре какого-либо слова, она считала каждую его форму отдельной леммой. Как видно в таблице, такие слова обычно попадают в один и тот же класс, например, *genoma* и *genomas* (*геном* и *геномы*), что говорит о правильной работе алгоритма классификации.

Также можно наблюдать «случайное размазывание» по разным классам массива наиболее общих терминов предметной области (*программа, компьютер, пользователь, файл, интерфейс*), что, вероятно, связано с их присутствием во многих текстах коллекции.

Как обычно бывает в случае автоматических методов обработки, результаты очень редко стопроцентно точны, скажем, в разные классы попали слова *информация* и *информационный*, *параллельный* и *параллелизм*, и пр.

Представленные термины относятся к разным областям информатики: биоинформатике, электронике, программированию, и пр. Заметим, что в нашей коллекции были тексты, скажем, из биоинформатики, поэтому и были термины из области биоинформатики. Это зависит от выбора текстовой коллекции, из которой извлекаются термины (Таблица 6).

Как обычно в случае онтологий, оценить полученные результаты непросто, потому что не существует «золотого стандарта» для оценки. Кроме того, ручное составление онтологии процесс очень субъективный, в этом смысле не очень ясно, может ли такой стандарт существовать. Особенно трудно оценить полноту (recall), т. е. насколько в построенной онтологии не хватает каких-либо терминов.

Наверное, если речь идет об оценке, то можно сравнивать с каким-либо словарем, это дало бы точность и полноту, в каком-то приближении, скажем, в нашем случае, со словарем по информатике. По-испански, у нас не было доступного словаря. Кроме

того, это не снимает вопроса о субъективности построения такого словаря, используемого для оценки.

В статье (Tingting He *et al.*, 2006) после ручной оценки были получены результаты с точностью в районе 70 %. Повторимся, что оценка ручная, потому что нет стандарта для оценки.

Если мы оценим точность выделения терминов нашим методом, то у нас получатся следующие результаты. Всего терминов 270, из них 31 глагол (подчеркнуты), то есть остается 239 терминов. Из этих терминов 19 явно не являются терминами данной области (зачеркнуты). Считая таким образом, мы получаем точность в 92,5 %. Если мы добавим к словам, которые будем считать неправильно определенными, 48 общенаучных терминов (курсив), то получим точность в [239 − (19 + 48)] / 239 = 72 %.

Табл. 6. Классы полученные в результате работы алгоритма

algoritmo, for, *implementación*, array, implementar, árbol алгоритм, for, *реализация*, массив, реализовать, дерево (поиска)
analógica, voltaje, binario аналоговый, напряжение, бинарный
as, if, int, integer, pseudocódigo, return, vtemp, *diagrama*, *descripción*, Turing, end as, if, int, integer (число), псевдокод, return (возврат), vtemp, *схема*, *описание*, Тьюринг, end (конец)
b2b, *business*, hosting, cliente, servidor, internet, ~~to~~, electrónico, consistir B2B, *бизнес*, хостинг, клиент, сервер, Интернет, ~~to~~, электронный, состоять
biología, bioinformática, adn, alineamiento, clustalw, fago, gen, genoma, genomas, genome, genómica, génica, homología, *human*, microarrays, *modelado*, nucleótidos, *predicción*, proteína, proteína-proteína, sanger, secuenciación, evolutivo, *secuencia*, *biológico*, computacional, protocolo, *variedad*, *análisis, técnica, estructura, interacción*, completar, *montaje, herramienta,* ~~menudo~~, usar, talar, software, visualizar, *cuantificación*, *modelo*, automatizar, búsqueda *биология*, биоинформатика, ДНК, выравнивание, ClustalW, фаг, ген, геном, геномы, геном, геномика, генный, гомология, *человек*, микрочип, *моделирование*, нуклеотиды, *прогнозирование*, белок, белок-белок, Sanger последовательность, эволюционный, *последовательность*, *биологический*, вычислительный, протокол, *отбор, анализ, технологический, структура, взаимодействие,* дополнить, *монтаж, инструмент,* ~~часто~~, ~~использовать~~, ~~рубить~~, программное обеспечение, визуализировать, *количественная оценка*, *модель*, автоматизировать, поиск
componente, transistor, tubo, funcionar, conexión, dispositivo, ~~etc~~, *tecnología*, digitales, microprocesadores, *velocidad, lógica*, soler, altavoz компонента, транзистор, трубка, функционировать, связь, устройство, ~~и т.~~ д., *технология*, цифровые, микропроцессоры, *скорость, логика*, случаться, динамик
computación, *ciencia, constable, científica,* cómputo, *disciplina, matemática*, ~~usualmente~~, *teoría*, computacionales, *ingeniería*, estudiar, artificial, *matemático*, informática, paralelo, programación компьютер, *наука, стабильный, научный,* вычисление*, дисциплина, математика*, ~~как правило~~, *теория*, вычислительные, *инженерный*, исследовать, искусственный, *математический*, информатика, параллельный, программирование
conjunto, *notación, problema*, finito, binaria, complejidad, np, np-completo, *número, tamaño, elemento*, coste, lineal, ~~comúnmente~~, ~~montículo~~ множество, *обозначение*, *проблема*, конечный, бинарный, сложность, NP, NP-полный, *количество, размер, элемент*, стоимость, линейный, ~~обычно~~, ~~курган~~
código, compilador, compiladores, lenguaje, máquina, programa, ~~compuesto~~ код, компилятор, компиляторы, язык, машина, программа, ~~состоящий~~
descifrar, criptografía, cifrar, *método*, texto, denominar декодировать, криптография, шифровать, *метод*, текст, обозначать
dimensión, cubo, *espacial*, almacén, marts, metadato, middleware, warehouse, data, olap, tabla, operacional, variable, *definición*, especificar, usuario, poseer, almacenar, dato, colección, arquitectura, registro *измерение*, куб, *пространственный*, хранилище (данных), marts, метаданные, промежуточное программное обеспечение, хранилище (данных), данные, OLAP, таблица, оперативный, переменная, *определение*, указать, пользователь, иметь, хранить, данные, набор, архитектура (компьютера), запись
diseñar, diseñador, objeto, funcional, procesar, proceso *разработка, разработчик*, объект, функциональный, обработать, процесс
formato, avi, compresión, *especificación*, formatos, mov, archivar, vídeo, audio, archivo, informático, codificar, *estándar* формат, AVI, сжатие, *спецификация*, форматы, MOV, архивировать, видео, аудио файл, информационный, шифровать, *стандартный*

potencia, válvula, analógicos, semiconductor, corriente, alternar, analizador, electrónica, conmutación, eléctrico, sonido, pila, supercomputadoras напряжение, клапан, аналоговые, полупроводниковый, ток, изменять (полярность), анализатор, электронный, коммутация, электрический, звук, аккумулятор, суперкомпьютеры
red, *principal*, *artículo*, permitir, utilizar, ~~vario~~, aplicación, información, ~~través~~, ~~tipo~~, *sistema*, ~~ejemplo~~, característica, interfaz, *forma*, gestión, operativo, acceder, ~~diferente~~, base, contener, operación, función, clasificar, ordenador, ejecutar, programador, cálculo, modelar, relacionales, interfaces, objeto, relacional сеть, *основной, статья,* позволять, использовать, ~~различный~~, приложение, информация, ~~посредством~~, ~~тип~~, *система*, ~~пример~~, характеристика, интерфейс, *форма*, управление, оперативный, обратиться (к данным), ~~разный~~, база, содержать, операция, функция, классифицировать, компьютер, исполнять (программу), программист, расчет, моделировать, реляционные, интерфейсы, объект, реляционный
~~rápido~~, acceso, ~~sencillo~~, soportar, web, ~~específico~~, ~~central~~, fiabilidad, paralelismo ~~быстрый~~, доступ, ~~простой~~, поддерживать, Web, ~~конкретный~~, ~~центральный~~, надежность, параллелизм
señal, transductores, transductor, impedancia, filtrar, conversión, acondicionamiento, convertidor, daq, *adquisición*, analógico, conectar, adaptación, frecuencia, medir, tensión, sensores, digital, cable, control, ~~física~~, entrada, medición, ~~físico~~, salida, ~~normalmente~~, bus, dato сигнал, датчики, датчик, сопротивление, фильтровать, преобразование, упаковка, конвертер, DAQ, *приобретение*, аналоговый, соединять, адаптация, частота, измерять, напряжение, датчики, цифровой, кабель, управление, ~~физика~~, ввод, измерение, ~~физический~~, выход, ~~как правило~~, шина (данных), данные
térmico, ci, cápsula, integration, scale, chip, circuito, chips, integrar, híbrido, silicio, reproductor, amplificador, ~~fabricación~~ тепловой, CI, капсула, интеграция, масштаб, микросхема, плата, микросхемы, комплексный, гибридный, кремний, проигрыватель, усилитель, ~~производство~~

4. Выводы

В данной статье мы представили метод, который позволяет построить проект онтологии предметной области состоящий их однословных терминов, используя тексты данной области и неспециализированный корпус. В дальнейшем эти термины можно объединять в многословные. Метод позволяет определить слова, которые являются возможными терминами (или частями многословного термина) данной области используя логарифмическую меру сходства. После этого дополнительно термины классифицируются на основе меры сходства по косинусу угла с использованием алгоритма классификации *k-means*.

Предварительная ручная оценка результатов работы метода показывает, что метод дает хорошие результаты. Полученный автоматически список терминов достатчоно велик и должен рассматриватся как первый шаг к построению бОльших списков. С другой стороны, количество терминов в текстах ограничено, то есть такой список не может расти бесконечно.

В качестве будущих направлений работы можно упомянуть следующие:

- Определить возможность автоматического определения порога при отборе терминов.
- Попробовать разные параметры алгоритма классификации *k-means*. Оценить возможность применения алгоритма, позволяющего определять количество классов автоматически.
- Вместо меры сходства по косинусу угла попробовать другие меры сходства при классификации.
- Сравнить разные логарфимические меры сходства при поиске терминов.
- Выполнить сравнение с одним или несколькими корпусами разных предметных областей, чтобы отфильтровать общенаучные термины.

Заметим, что имплементация метода не представляет каких-либо существенных трудностей.

Литература

1. *Caraball S. A.* Automatic construction of a hypernym-labeled noun hierarchy from text. // In: Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, 1999.
2. *Cimiano P.* Ontology learning and population from text, algorithms, evaluation and applications. // New York, USA: Springer, 2006.
3. *Dunning T.* Accurate methods for the statistics of surprise and coincidence. // Computational Linguistics 19.1 (Mar. 1993), 61–74.
4. *Gelbukh A., Sidorov G.* On Automatic Morphological Analysis of Inflective Languages. // In: Proc. of International Conference on computational linguistics and its applications Dialogue-2005 (in Russian), 2005, Russia, pp 92–96.
5. *Gómez-Pérez A., Fernandez-López M. & Corcho O.* Ontological Engineering. // London: Springer Verlag, 2004.
6. *Maedche A., Staab S.* Discovering conceptual relations from text. // In: Proceedings of ECAI 2000, 2000.
7. *Punuru J.* Knowledge-based methods for automatic extraction of domain-specific ontologies. // PhD thesis, 2007.
8. *Rayson P., Berridge D. and Francis B.* Extending the Cochran rule for the comparison of word frequencies between corpora. // In: Volume II of Purnelle G., Fairon C., Dister A. (eds.) Le poids des mots: Proceedings of the 7th International Conference on Statistical analysis of textual data (JADT 2004), Louvain-la-Neuve, Belgium, March 10–12, 2004, Presses universitaires de Louvain, pp. 926–936.
9. *Tingting He, Xiaopeng Zhang, Ye Xinghuo.* An Approach to Automatically Constructing Domain Ontology. // PACLIC 2006, Wuhan, China, 1–3 November, 2006, pp. 150–157.
10. *Uschold M. & Grunninger M.* Ontologies: Principles Methods and Applications. // Knowledge Egineering Review, 1996.

Такая девочка-девочка. Семантика редупликации существительных в русской разговорной речи и языке интернета[1]

Such a girl-girl. Semantics of noun reduplication in colloquial russian and the internet language

Гилярова К. А. (hilaris@yandex.ru)

Институт лингвистики РГГУ, Москва

В работе рассматривается редупликация существительных в разговорном русском языке, в таких примерах как «такая девочка-девочка», «прям лето-лето» и др. Семантический анализ показывает, что удвоенные существительные могут нести значение прототипа, коннотации, уточнения, детерминатива, усиления и положительной оценки.

Введение

Рассмотрим следующие примеры, взятые из блогов (1), радио-эфира (2) и разговора по ICQ (3):

(1) *И такая* ***осень-осень, Питер-Питер*** *под окнами*[2].

(2) *В кино Андрей Сергеевич совсем не такой, как в театре, потому что там он совсем уже* ***мэтр-мэтр*** (А. Домогаров об Андрее Кончаловском в передаче «Дифирамб» на «Эхе Москвы»)

(3) *— А чем они занимаются? — Да я уже и забыла, но совсем не детьми и не* ***наукой-наукой****, а скорее какой-то историко-правозащитой.*

Подобные удвоения существительных все чаще встречаются в разговорной речи молодого поколения (в основном школьного и студенческого возраста) и во всемирной сети Интернет. Они и стали объектом настоящего исследования.

Остановимся подробнее на следующих примерах:

(4а) *Сол тоже приезжает туда, они снимают номер в мотеле и…* ***любовь, любовь, любовь****.* [Нина Катерли. Дневник сломанной куклы // «Звезда», 2001]

(4б) *Если у них такая* ***любовь-любовь*** *в 14, пусть женятся.*

(5а) *Ой,* ***мороз, мороз,*** *не морозь меня.*

(5б) *Ну, сейчас на улице −18 и я бы не сказал, что прям* ***мороз-мороз.***

В примерах (б) удвоенное существительное произносится без паузы и представляет собой целостную просодическую единицу, в то время как в (а) запятая указывает на наличие паузы. Вслед за А. Вежбицкой [Вежбицкая 1999], **редупликацией** (или удвоением) мы будем называть лишь феномен, представленный в (4б) и (5б). А явление, отраженное в (4а) и (5а) назовем **повтором** и оставим за рамками настоящего исследования. Речь пойдет о редупликации **синтаксической**: мы будем анализировать лишь не закрепленные в словарях удвоения, и удвоения слов, а не морфем. Наконец, в рамках данной работы не представляется возможным описать семантику редупликации всех частей речи, и мы ограничили материал существительными[3]. Такой выбор обусловлен новизной явления. Еще не-

[1] Автор благодарит за помощь в подготовке доклада Л. Д. Беклемишева, В. И. Киммельмана, Т. Э. Русситу.

[2] Во всех приведенных в работе примерах исправлены орфографические и пунктуационные ошибки, но больше никаких изменений не внесено: оставлены порядок слов, авторская пунктуация, разговорная лексика, «падонкаффский» жаргон.

[3] О редупликации прилагательных, наречий, местоимений и местоименных наречий см. [Крючкова 2004], о редупликации глаголов — [Israeli 1997].

давно полное удвоение существительных в русском языке считалось невозможным, и сейчас оно все еще не зафиксировано в грамматиках, по крайней мере, далеко не в том объеме, в каком встречается в современном разговорном языке.

Согласно сравнительно недавним исследованиям, для существительных в русском языке характерно дивергентное удвоение, часто с заменой начального согласного первого компонента губно-губным согласным в составе второго компонента. Ср. *тарыбары, фигли-мигли, фокус-покус, шалтай-болтай, шуры-муры, танцы-шманцы, мастер-ломастер, страсти-мордасти, трава-мурава, коза-дереза* и др. [Беликов 1990], [Земская и др. 1981], [Израэли 1996], [Крючкова 2004], [Plähn 1987]. Упоминаний полной редупликации существительных в литературе немного. Зачастую высказывания, приводимые как примеры удвоения, на самом деле имеют посередине ощутимую паузу и потому тяготеют к клаузальному или внутриклаузальному повтору[4] (Ср. 1а, 2а). Однако А. Израэли обращает внимание на случай, когда выражение «X-X, а…», где X — существительное в именительном падеже, означает ‘несмотря на то, что субъект X, ему присущи какие-то не характерные для X-а свойства’ [Israeli 1997, с.592]:

(6) *Видят они:* ***дурак-дурак****, а не дурее других.*

Л. Л. Федорова пишет о редупликации существительных в фольклорных текстах и сказках, особенно в обращениях (*Я колобок-колобок, Петушок-петушок!* и пр.), однако справедливо указывает на ограниченность этой модели [Федорова 2005, с. 197].

До последнего времени казалось, что существительные в русском языке не могут выстраивать продуктивные модели удвоения. Зато подобные конструкции уже по меньшей мере полвека назад вошли в английский язык. В [Ghomeshi др. 2004] коллектив авторов подробно рассматривает примеры вроде:

(7) *It's tuna salad, not* ***SALAD-salad.***

(8) *Do you* ***LIKE-HIM-like him****?*

Такие модели авторы называют «сопоставительной редупликацией» или CP (“contrastive reduplication”, CR). CP распространена в современном английском, встречается в речи как молодых, так и пожилых носителей языка и охватывает не только слова разных частей речи вплоть до предлогов, но и целые выражения. Л. Хорн, в свою очередь, называет CP «лексическим клонированием» (“lexical cloning”) и выделяет четыре ее семантических типа: 1) прототип, 2) буквальное значение, 3) усиление и 4) добавочное значение [Ghomeshi и др. 2004, с.314].

Синтаксическая редупликация, в том числе существительных, встречается также во французском, итальянском [Вежбицкая 1999], [Федорова 2005], испанском [Horn 1993] языках. Ср. [Федорова 2005, с.197-198]:

(9) *Siamo* ***amici-amici****? — Мы настоящие друзья?* (букв. «друзья-друзья»)

(10) *Abiti a* ***Mosca-Mosca****? — Ты живешь в самой Москве?* (т. е. не в пригороде)

В настоящей работе проводится семантический анализ аналогичных конструкций в русском языке.

1. Методы исследования

Для создания базы примеров мы пользовались системой поиска по Национальному корпусу русского языка (www.ruscorpora.ru) и «точечным» поиском в системе Яндекс. Список слов для точечного поиска сформировался в результате анализа примеров, услышанных в разговорной речи, по радио и телевидению, а также случайно замеченных в сети. Всего было проанализировано более 400 примеров.

Удвоенные существительные встречаются почти исключительно в предикативной функции и зачастую предваряются усилителями *такой*, *прямо* (с разговорным вариантом *прям*), *совсем уж* или входят в противопоставительные конструкции *не X-X, а…* или же *не…, а X-X*. На письме сдвоенное слово пишется чаще всего через дефис, реже — через пробел.

Часто удвоенные существительные встречаются в комментариях к фотографиям. Как правило, такие примеры содержат только голый повтор, безо всякого контекста, и потому не дают достаточно материала для семантического анализа. Зато отличным подспорьем являются сами фотографии: по ним обычно можно с точностью определить, что именно вызвало у комментирующего потребность употребить редупликацию. Чтобы выяснить, существуют ли какие-то общие закономерности в выборе удвоенной формы, или же пока носители пользуются редупликацией лишь в рамках своего идиолекта, мы провели небольшой эксперимент. Фотографии, найденные в сети по ключевым словам с удвоением (*девочка-девочка, бабушка-бабушка, зима-зима* и пр.), были предъявлены 27 респондентам (средний возраст — 24 года) с просьбой описать картинку одним редуплицированным словом. Для 29 картинок из 33 нашелся хотя бы один респондент, использовавший именно ключевое слово, по которому фотография была найдена. Для 20 картинок из этих 29

[4] Термины «клаузальный» и «внутриклаузальный повтор» вводит А. Вежбицкая [Вежбицкая 1999, с.229, 240]

ключевое слово было самым распространенным вариантом ответа. Так, например, изображение Сары Мишель Геллар (исполнительницы главной роли в сериале «Баффи») в коротких шортах и розовом топе 13 респондентов (50%) прокомментировали как *блондинка-блондинка*, а фотографию целующейся пары подростков 9 человек (30%) описало как *любовь-любовь* и еще двое как *любовь-морковь* (другие сравнительно популярные ответы — *поцелуй-поцелуй* и *молодость-молодость*). Про заснеженную сосновую ветвь 16 респондентов (59%) написали *зима-зима*, а про желтые кленовые листья — *осень-осень*. Все эти ответы культурно обусловлены, так что редупликация может считаться еще одним средством, помогающим выявить прототипы и коннотации.

2. Семантика редупликации существительных в русском языке

Согласно нашему исследованию, удвоение существительных в разговорной речи может иметь следующие значения.

2.1. Прототип. X-X = ‘прототипический, обладающий характерными признаками X-а X’

(11) *У вас прямо **свадьба-свадьба** была, или вы ограничились регистрацией и узким семейным кругом?*

(12) *Знакома с семьей из папы-мамы-2х детей, тоже милые, на великах-роликах вместе катаются, ну прям **семья-семья**, образцово-показательная.*

(13) *Я сняла дачу — такую **дачу-дачу**… с местом для шашлыка и вечернего чая, с видом на закат…*

(14) *Все женщины разные, и по жизненному опыту знаю, что далеко не все мамы прям «**мама мама**»! Мое окружение было на все 100 уверенно, что я на 5 день после родов побегу на переговоры и т. д. А я оказалась «сильно **мама мама**».*

Под «X-X» в (11)-(14) скрывается прототип X-а [Rosch 1975]. Прототипическая в нашей культуре свадьба подразумевает белое платье, лимузин, шампанское, обручальные кольца, застолье с гостями. Прототипическая семья — это полная и счастливая семья с двумя детьми. Обязательные атрибуты настоящей дачи — сад, шашлык, а также, судя по анализу фотоматериала, тюлевые занавески, старый транзистор и букет полевых цветов на круглом столе с белой скатертью. Прототипическая мама сама заботится о своем ребенке и некоторое время после родов не выходит на работу. Характерно, что удвоения, выражающие прототип, понимаются слушающим безо всякого поясняющего контекста:

(15) *Очень хорошие фотографии и рассказ, такое **лето-лето**!*

Анализируя корпус примеров на редупликацию существительных, можно создавать целые портреты, соответствующие народному представлению о девочке, бабушке, маме, мальчике, «блондинке», женщине, мачо, боссе т. д. Так, например, если собрать контексты из всех найденных примеров, содержащих пару *девочка-девочка*, получится, что прототипическая девочка:

носит бантики, косички, яркие платья, сарафанчики в рюшечках, футболки в оборочках, чешки; с утра до ночи перекрашивает ногти, красит губы; имеет огромные глазищи кружком и длинные ресницы; любит пуфики и рюшечки, покупки и пирожные, розовое и голубое; не делает ничего сама; не разбирается в компьютерах; не говорит и не пишет ничего циничного и грубого; наконец, когда есть лишние полчаса, идет по магазинам.

А *мужиком-мужиком* можно назвать:

взрослого мужчину лет 40–50 с пробивающейся сединой или лысиной и другими признаками преклонного возраста; который носит кожаную куртку, кожаную кепку, спортивную сумку, допотопную «мобилку», клетчатую рубашку; с 13 лет работает, заботится о маме и младшей сестре; лампочку прикрутит, всё починит; настоящий глава семьи; мужественный, уверенный в себе мужчина, такой, что в его объятьях можно утонуть, рядом с ним чувствуешь себя женщиной, хочется лечь и лапки поднять.

Обращение к прототипу — наиболее часто встречающееся значение редупликации.

2.2. Коннотация. X-X = ‘Y, обладающий некоторыми прототипическими свойствами X-а’

(16) *Вафелька вся такая мимими… Лизучая, игручая и прям вот такая **девочка-девочка.** <…> Целует в нос, хрюкает, поскуливает…* (о собаке)

(17) *Я считаю, <…> человек, стоящий на входе в подобные места, должен знать, кого именно он не пускает. Ужасно неприятно, такая **Москва-Москва*** (речь идет о том, что кого-то не пустили в московский ночной клуб в неподходящей одежде).

(18) *Я бы покрасила и поцарапала потом. Чтобы была такая* ***деревня-деревня****, французский антик* (про самодельный диванчик, у которого спинкой служит старая дверь).

Судя по корпусу примеров и фотоматериалу, выражение «девочка-девочка» зачастую применяется по отношению к взрослым женщинам и даже кошкам и собакам (ср. 16). В таком случае оно означает некий набор признаков, характерных для образа прототипической девочки, но уже оторвавшихся от него и заживших собственной жизнью, ставших коннотацией. Согласно Ю. Д. Апресяну, коннотации — это «несущественные, но устойчивые признаки выражаемого лексемой понятия, которые воплощают принятую в данном языковом коллективе оценку соответствующего предмета или факта действительности» [Апресян 1995, с.159]. Коннотации «не входят непосредственно в лексическое значение слова и не являются следствиями или выводами из него» [с.159]. Представления о прототипической свадьбе не входят в толкование, но выводятся из него, ср. свадьба = ‘обряд заключения брака, а также празднество по случаю вступления в брак’ [МАС]. Представления о классической девочке никак не следуют из значения лексемы, ср. девочка = ‘ребенок или подросток женского пола’ [МАС].

В (17) говорящий апеллирует к представлению о Москве как о недружелюбном закрытом обществе, а о москвичах как о заносчивых, не внимательных к окружающим, судящих «по одежке» людях. Подобная коннотация столицы уже устоялась настолько, что *Москва-Москва* встречается в таком значении и безо всякого поясняющего контекста:

(19) *Я не могу больше в симачев ходить, там сейчас такая* ***москва-москва****…* («симачев» — московский ночной клуб-бар «Денис Симачев»)).

Предъявленную респондентам фотографию глухой подворотни, найденную в сети по ключевому слову *Питер-Питер*, несколько москвичей и одна киевлянка описали как *питер-питер*, в то время как участвовавшие в опросе петербуржцы все написали *колодец-колодец* или *двор-двор*. Выражение *Питер-Питер* отсылает к некому представлению иногородних о городе на Неве, представлению, не имеющему ничего общего с нашими энциклопедическими знаниями о Санкт-Петербурге и тем более с восприятием Петербурга его коренными жителями.

Как отличить протитип Х-а от коннотации Х-а? Можно с уверенностью говорить о коннотации, если удвоенное существительное употреблено при «неподходящем» референте. Т. е., например, «девочкой-девочкой» названа бабушка, «мамой-мамой» — старшая сестра, «учительницей-учительницей» — строгая женщина с пучком и в очках и т. д. Редуплицированные существительные в таком случае имеют уже совершенно иное значение, нежели их «одинарные собратья». Так, выражение *москва-москва* можно истолковать как ‘неприветливый’, а *девочка-девочка* как ‘мечтательная’, ‘любящая наряжаться’, ‘глупая’. В следующем примере редупликация порождает значение ‘белый цвет’ у слова *зима* (что подтверждается фотографией):

(20) *В свитере получилась хорошая сессия. На выкладке цвета про…лись, но если смотреть в полный размер, то прям* ***зима-зима****.*

2.3. Точность. Х-Х= ‘точно Х, в буквальном смысле Х, именно Х и ничто другое’

(21) *Нет ли у кого на примете горнолыжного инструктора в Киеве? Не обязательно прям* ***инструктора-инструктора****, но чтобы человек сам хорошо и технично катался + умел рассказать-показать.*

(22) *Ну, не прям* ***школа-школа*** *— пока только по три урока по субботам.*

(23) *Блин, жалко, что тебе прям* ***море-море*** *=)) я сча тоже в отпуск собираюсь… я правда в Чехию и Литву планирую…*

(24) *Я не понял, так мы про* ***секс-секс*** *или про жОткий петтинг?*

Может показаться, что в (21)–(24) речь идет о прототипе, как в (11)–(15). Но это не так. Свадьба (11) не перестает быть свадьбой, даже если вовсе не отпразднована, семья (12) по-прежнему является семьей, даже если далека от прототипического варианта «мама, папа и двое детей». А вот умения технично кататься на лыжах и хорошо объяснять, строго говоря, недостаточно, чтобы считаться инструктором, как и три урока по субботам не могут считаться школой.

Во (21)–(24) удвоение лексемы отсылает к ее основному значению, подчеркивает, что слово употреблено в буквальном смысле. *Инструктор-инструктор* — это инструктор по профессии и только он, *школа-школа* — регулярные уроки в настоящей школе, *море-море* — именно море, а не озеро и даже не залив, под *сексом-сексом* подразумевается только половой акт. Редупликация устанавливает для Х-а строгие рамки: или Х, или не Х, а третьего, промежуточного варианта нет. «Повторяя слово (‘Х-Х’), говорящий привлекает внимание к этому слову и настаивает на его строгом соответствии требованиям истины. ’Я имею в виду Х, а не что-то немного другое, чем Х’» [Вежбицкая 1999, с. 236] Согласно А. Вежбицкой, с помощью этого определения можно полностью описать семантику синтаксси-

ческой редупликации в итальянском. Однако лишь малую толику примеров из нашего корпуса удалось истолковать таким образом.

Чтобы проверить в каждом конкретном случае, несет ли в себе редупликация значение прототипа или уточняет буквальное значение слова, можно заменить удвоенную единицу ее «одинарным» аналогом. Если повтор лишь подчеркивал буквальное значение, общий смысл фразы не изменится, ср.:

(22а) *Ну, не прям **школа** — пока только по три урока по субботам.*

Если же редупликация отсылала к прототипу или коннотации, фраза станет бессмысленной:

(12а) **Знакома с семьей из папы-мамы-2х детей, тоже милые, на великах-роликах вместе катаются, ну прям **семья**, образцово-показательная.*

(15а) **Я не могу больше в симачев ходить, там сейчас такая **москва** …*

2.4. Интенсивность. X-X =’высокая степень X-а’

Во всех следующих примерах редупликация несет усилительное значение:

(25) *Во-вторых, вот как раз в ноябре там уже была не прям **жара-жара**…* (= ‘сильная жара’)

(26) *Я еще повспоминала… все-таки ощущение свободы в детстве — это было ого-го, вот прямо небо на плечах, прямо **счастье-счастье*** (=’большое счастье’).

(27) *За окном такой дождь, прямо **дождь-дождь!*** (=’сильный дождь’)

(28) (Проехала пожарная машина) — *Смотри, где-то пожар.* (Проехала еще одна) — *О, прям **пожар-пожар!*** (=’сильный пожар’)

(29) *Еще такое **утро-утро**, а я на работе* (=’раннее утро’).

Такого значения редупликации естественно было ожидать, ведь удвоение прилагательных и наречий (*большой-большой, быстро-быстро*), давно узаконенное в грамматиках и являющееся нормой для литературного русского языка, выражает именно интенсивность (признака или действия) [Крючкова 2004]. Редупликации подвергаются, как правило, качественные прилагательные (ср. **деревянный-деревянный*). Так и круг существительных, способных порождать удвоения со значением «интенсивность», ограничен лишь теми, денотаты которых можно измерить.

Возникает вопрос, стоит ли выделять интенсивность в отдельное значение, ведь можно считать, что *жара-жара* — это прототипическая жара, *счастье-счастье* — настоящее счастье и т. д. Однако высокая степень X-а не всегда означает его прототипичность. Так, сильнейший пожар на газопроводе в Москве в мае 2009 года со столбом огня в 200 метров вряд ли можно считать прототипическим. Сомнительно также, что сильный короткий ливень следует считать прототипом дождя (скорее уж это необходимый атрибут прототипической грозы). Регулярность «усилительного» удвоения прилагательных и наречий также говорит в пользу выделения усилительной функции и у редупликации существительных.

Наконец, по нашим наблюдениям, в целях передачи интенсивности существительные могут не только удваиваться, но и утраиваться: *жара-жара-жара* = ‘особо сильная жара’, *любовь-любовь-любовь* = ‘очень сильная любовь’. А *ужас-ужас-ужас* даже легло в основу анекдота и стало использоваться как идиома. Для выражения прототипического и буквального значений утроение никогда не используется. Ср. **такая девочка-девочка-девочка*, **прям инструктор-инструктор-инструктор*.

В [Земская и др. 1981] как сильное экспрессивное средство отмечается повтор с префиксацией. «Второй член повтора содержит приставку *пере-, рас-* или *раз-*, придающую усилительное значение всей конструкции» [с.115]:

(30) *Он **москвич-размосквич** в пятом поколении.*

(31) *Мы такие с вами **соседи-рассоседи**.*

(32) *Здесь он **академик-переакадемик**, а в Москве был бы просто доцентом.*

Судя по тому, что молодое поколение носителей не только не употребляет, но и не всегда понимает и признает такие конструкции, они постепенно уходят из языка, оставляя место полным удвоениям — коротким, минималистским, и, возможно, скопированным с языков с менее развитой морфологией и словообразованием. «Друг-друг» встречается в поисковой системе Яндекс в несколько раз чаще, чем приводимый в [Земская и др. 1981] «друг-раздруг».

2.5. Положительная оценка. X-X = ‘хороший X’, ‘стоящий X’, ‘красивый X’

(33) — *И как она тебе? — Ну, там прямо такие **ноги-ноги**…* (= ‘красивые, стройные, длинные ноги’)

(34) *Ой, ну уж прямо изображая жертву прямо* ***фильм-фильм*** (в ответ на внесение фильма «Изображая жертву» в топ лучших фильмов).

(35) *Лентус… да я за половину этой цены такое* ***платье платье*** *Полинке сошью… что и ты, и она до своих пенсий этой фоткой всем хвастаться будете.*

(36) *Спасибо, что есть такая* ***песня-песня****, наполняющая душу воспоминаниями молодости* (о песне Визбора «Люди идут по свету»).

В (33)-(36) повторение существительного добавляет высказыванию оценочный компонент, причем всегда положительный. *Песня-песня* в (36) не может быть заменена на просто *песня* без ущерба для смысла. Для равносильной замены следует выбрать *хорошая песня*. А фраза (34) без удвоения и вовсе бессмысленна.

2.6. Разрешение многозначности. X-X='X в определенном значении'

(37) *Чем, по-вашему, отличается парень-друг от* ***парня-парня****? Я уже перестала понимать, что к чему. Особенно сложно это разбирать в самом начале отношений.*

(38) *Ищу* ***девушку-девушку****. Работу не предлагать!*

(39) *При просмотре Хауса я все время не могу отделаться от ощущения, что там какой-то конец восьмидесятых: мало того, что сообщения приходят на пейджер (правда, мне уже умные люди объяснили, что это не* ***пейджер-пейджер****, а внутрибольничная автоматизированная система рассылки оповещений, так что ладно) <…>*

В (37)–(39) удвоение существительного отсылает к определенному значению многозначного слова. Так, парнем можно назвать как любого юношу и даже вообще любого мужчину (ср. парень 1, парень 2 в [МАС]), так и любимого человека, «бой-френда». Сдвоенное употребление *парень-парень* указывает именно на последнее значение.

Семантику редупликации в 2.6 можно считать частным случаем 2.3. В самом деле, в обоих случаях речь идет об уточнении значения слова, говорящий с помощью удвоения подчеркивает, что именно хочет сказать. Тем не менее, мы разделили 2.3 и 2.6. В 2.3 редупликация означает 'именно X, а не не-X', а в 2.6 — 'именно это значение X из нескольких'.

2.7. Детерминатив. X-X='определенный X'

Удвоение существительных может нести функцию определенного артикля, помогая слушающему выбрать конкретного референта из многих.

(40) — *Вот я с Сашами* (на фотографии) — *Ну и который из них* ***Саша-Саша****?* (то есть 'тот самый Саша', 'Саша, о котором шла речь раньше, или с которым у говорящего особые отношения').

Иногородние студенты говорят как о *доме-доме* о родительском доме в родном городе, в то время как их московское жилище для них просто *дом*. Студенты 2-го курса Института лингвистики РГГУ называют *буфетом-буфетом* буфет в их учебном корпусе, а буфет главного корпуса, соединенный со столовой, носит у них имя *буфет-столовая*. Про все остальные буфеты они говорят нейтрально *буфет*. Эти же студенты рассказали, что называют *кино-кино* их любимый кинотеатр «5 Звезд», куда они ходят чаще всего:

(41) *А в какое кино пойдем?* ***Кино-кино****?*

2.8. Слияние значений

Для некоторых лексем трудно провести границу между разными значениями. Так, представление о прототипическом друге неотделимо от понятия «хороший друг», поэтому следующие примеры можно отнести сразу к 2.1 и 2.5:

(42) *Лёха уезжает. Грустно… Он такой* ***друг-друг.***

(43) *Саш, ты прелесть… ты прям такой* ***друг-друг****… поддержишь всегда, выслушаешь…..спасибо тебе))*

(44) *А Алина… да ну ее нафиг, честно. Вроде такой* ***друг-друг****, а сама ни разу не позвонила.*

Аналогично настоящая любовь — это сильная любовь, так что удвоение *любовь-любовь* соединяет в себе значения 2.1 и 2.4. Ср. (4б), а также:

(45) *Лето уже хочу, на мопедике по Маросейке и влюбиться бы уже взаимно, чтоб прям* ***любовь-любовь****)))*

2.9. Многозначность редупликации

Редупликация одного и того же слова может нести разные значения в зависимости от контекста. Так, например, *няня-няня* может означать: — прототипическую няню (любящая, простая, заботится

о детях) (2.1); — коннотацию няни (носит чепчик, вяжет, поет колыбельные, необразованна) (2.2); — профессиональную няню из агентства (ср. 2.3); — хорошую няню, супер-няню (2.5); — конкретную няню на фоне многих или в противопоставлении другой няне (2.7).

В (46) и (47) у выражения *ремонт-ремонт* совершенно разные значения: в (46) ‘хорошо сделанный ремонт’, то есть уже результат (2.5), а в (47) — ‘высокая степень ремонта’ (2.4) или ‘настоящий ремонт’ (2.1):

(46) *Зато я еще позавчера сидела на втором этаже, где у нас — ну прямо* ***ремонт-ремонт****, курила и смотрела концерт Богушевской.*

(47) *Там не такой уж прям* ***ремонт-ремонт****, чтобы нас выселяли; так, косметический…*

Аналогично *в (36)* ***песня-песня*** интерпретируется как ‘хорошая, стоящая песня’, а в (48) — *как* прототипическая, настоящая песня со словами и музыкой:

(48) *Вчера написала свою первую песню. Ну то есть попытки всякие были и раньше, но чтобы вот так получилась прямо* ***песня-песня****, со своей оригинальной мелодией и словами, — это впервые.*

Заключение

В работе рассмотрена семантика удвоения существительных в разговорной речи и языке Интернета. Анализ собранных данных показывает, что редупликация существительных *X-X* в русском языке может выражать 1) прототип X-а, 2) коннотацию X-а, 3) точное, буквальное значение X-а, 4) усиление X-а, 5) положительную оценку X-а, 6) уточнение значения многозначного X-а, 7) определенный референциальный статус X-а.

В дальнейшем интересно было бы исследовать, какого рода существительные прежде всего подвергаются удвоению, каковы синтаксические особенности таких конструкций и выделяются ли найденные нами значения и у редупликации других частей речи, также распространенной в современном русском языке. Проведению подобного исследования очень способствовала бы программа, позволяющая искать в поисковых системах Интернета любые повторы.

Литература

1. *Апресян Ю. Д.* Интегральное описание языка и системная лексикография. М.: Языки русской культуры, 1995. Т. 2.
2. *Беликов В. И.* Продуктивная модель повтора в русском языке // Russian Linguistics, 1990. Т. 14. С.81–86.
3. *Вежбицкая А.* Редупликация в итальянском языке: кросс-культурная прагматика и иллокутивная семантика // Семантические универсалии и описание языков (пер. с англ.). М.: Языки русской культуры, 1999. С.224–259.
4. *Земская Е. А., Китайгородская М. В., Ширяев Е. Н.* Русская разговорная речь. М.: Наука, 1981.
5. *Исраэли А.* Лексическая редупликация в русском языке // Russian Language Journal, 1996. № 165–167. С. 83–90.
6. *Крючкова О. Ю.* Вопросы лингвистической трактовки лексической редупликации в русском языке // Русский язык в научном освещении. М.: Языки славянской культуры, 2004. № 2(8). С. 63–85
7. *МАС* — Словарь русского языка в четырех томах. // Под ред. А. П. Евгеньевой. 2-е изд., испр. и доп. М.: Русский язык, 1981–1984.
8. *Федорова Л. Л.* Эмоции в грамматике // Эмоции в языке и речи. М.: Изд-во РГГУ, 2005. С. 178–199.
9. *Ghomeshi J., Jackendoff R., Rosen N., Russell K.* Contrastive focus reduplication in English (the Salad-Salad paper) // Natural Language & Linguistic Theory, 2004. V.22. P. 307–357.
10. *Horn L.* Economy and Redundancy in a Dualistic Model of Natural Language // S. Shore and M. Vilkuna (eds.) Yearbook of the Linguistic Association of Finland. SKY1993. P.31–72.
11. *Israeli A.* Syntactic reduplication in Russian: a cooperative principle device in dialogues // Journal of Pragmatics, 1997. V. 27. P. 587–609.
12. *Plähn J.* Хуйня-муйня и тому подобное // Russian Linguistics, 1987. Т. 11. С. 37–41.
13. *Rosch E.* Cognitive Reference Points // Cognitive Psychology, 1975. № 7. P. 532–547.

Концептуальные переменные образа мира по данным ассоциативных словарей

Conceptual variables of the world image according to the data of the associative dictionaries

Гольдин В. Е. (goldinve@yandex.ru)

Саратовский государственный университет им. Н. Г. Чернышевского

В докладе данные ассоциативных словарей использованы для установления комплекса основных характеристик (свойств, качеств, параметров), из которых в сознании человека складывается образ мира.

1. Введение

Конечная цель осуществляемого исследования состоит в том, чтобы, используя материал ассоциативных словарей, установить предсказуемые направления ассоциативных связей между стимулами и реакциями и таким путем приблизиться к пониманию принципов организации образа мира в сознании носителей языка.

Знание этих принципов может представлять интерес для создателей семантических и идеографических словарей (возможно, — и для разработчиков онтологий), поскольку в этих областях научных и прикладных исследований решаются задачи сходного типа и разбиению подвергаются множества соотносимых между собой сущностей: семантические словари устанавливают систему значений, закрепленных в лексике языка, идеографические — систему понятий, получающих вербальную экспликацию, в ассоциативном словаре мы ищем комплексы вербализованных ассоциативных связей как форм осознания мира.

Реализуемый в работе подход к установлению комплексов вербальных ассоциаций соотносим также по ряду параметров с поисками «естественного» состава тематических групп слов, с изучением единиц знания, с исследованием строения информационного тезауруса человека; при этом соотносимость не означает тождества, и имеющиеся различия достаточно существенны.

2. Исходные положения

1. Ассоциативные связи, реализуемые в свободном ассоциативном эксперименте, обнаруживают зависимость от грамматического класса стимулов: ассоциативные поля тех стимулов, которые в предъявляемой испытуемым форме могут восприниматься как законченное однословное высказывание, называющее фрагмент мира (*слон, окно, ураган, предложение* и под.), отличаются от ассоциативных полей тех стимулов, которые в предъявляемой испытуемым форме не могут быть восприняты как законченное однословное высказывание, называющее фрагмент мира (например, *телефона, дождю, переходить, разглядывая, новый, кувырком, вблизи* и под.). Следовательно, ассоциативные поля стимулов первого и второго типа должны рассматриваться раздельно. В настоящей работе рассматриваются ассоциативные поля стимулов-существительных в их начальной форме.

2. Известно, что направления развертывания ассоциативных полей стимулов-существительных коррелируют с семантическими типами этих существительных; при этом наиболее существенным оказывается различие между именами с предметным значением (имена естественных классов, артефактов, событий), с одной стороны, и именами с характеризующим значением (*умница, разгильдяй* и под.), с другой стороны. В настоящей работе рассматриваются ассоциативные поля стимулов-существительных с предметным значением.

3. Ассоциативные словари (в отличие от словарей толковых) не формулируют значений слов и не дают достаточных оснований для таких формулировок, но позволяют устанавливать с ф е р ы р е ф е р е н ц и и слов-стимулов, то есть определять те фрагменты мира, с которыми соотносится стимул в сознании испытуемых (например, на стимул стиль школьники и студенты дают большое количество реакций типа *одежда, одеваться, мода, модный, имидж, красота, фасон, джинсы* и под., что

определенно указывает на сферу моды как на одну из сфер референции данного стимула, но этого недостаточно для формулировки конкретного значения слова «стиль»). Считаем, что по крайней мере часть направлений ассоциативного развертывания поля может определяться через указание типичных для стимула сфер референции.

4. Ассоциативные реакции отражают ту меру подробности картины мира, ту степень системности образов мира и то сочетание «наивных» и научных представлений о мире, которые в действительности существуют в сознании испытуемых и к овнешнению которых вынуждает испытуемых процедура эксперимента. Именно эта картина мира рассматривается в данном случае как естественная.

5. Одна из главных функций сохранения в памяти образов мира (возможно, самая главная) — обеспечивать ориентацию людей в ситуациях, составляющих среду их существования. Ситуации же (и события как форма перехода одних ситуаций в другие), имеют композиционный характер: складываясь из предметов, качеств, состояний, действий, связей, отношений, аспектов, мотивов, целей, их различных сторон и т. п., ситуации в то же время характеризуются целостностью, и поэтому задача поиска типичных ассоциативных комплексов (и в результате — принципов организации образа мира в сознании носителей языка) представляется достаточно осмысленной.

6. В многочисленных работах исследователей (А. Р. Лурия, Н. А. Гасица, А. А. Залевская, Н. В. Уфимцева, Р. М. Фрумкина, Е. И. Горошко, Ю. Н. Караулов, Г. А. Мартинович и др.) установлены варианты ассоциативных реакций, подробно классифицированы типы отношений между стимулами и реакциями, показано, что основные из этих типов фиксируются во всех ассоциативных полях (Гасица: 10, 11); психолингвистикой подготовлен и следующий шаг: представить разновидности ассоциаций в качестве взаимосвязанных сторон единого комплекса знаний, обеспечивающих ориентацию в мире.

3. Исследование

Материалом, рассматриваемым в работе, являются данные двух ассоциативных словарей: Русского ассоциативного словаря [РАС 2002] и Русского ассоциативного словаря школьников (см.: [АСШС 2009]).

Основной метод решения поставленной задачи — итеративное заполнение фреймов исследуемых ассоциативных полей, в ходе которого включение нового материала в одних случаях диктует более дробную организацию схемы, а в других открывает возможность укрупнения уже установленных в ней слотов. Итеративность заключается в том, что фрейм, построенный на материале ассоциативных полей одной части отобранных для исследования стимулов, подвергается коррекции с учетом ассоциаций на стимулы другой части корпуса стимулов, а затем исследователь проверяет адекватность скорректированной схемы, вновь заполняя ее материалом тех ассоциативных полей, с которых начиналось исследование.

Например, в ряду ассоциаций на стимулы-существительные, являющиеся номинациями диких животных (лиса, волк, медведь и др.), обнаруживаются реакции типа А1 (*нора, берлога* и под.), типа Б1 (*лес, поле, тайга* и под.), типа В1 (*Африка, Австралия*), а также типа Г1 (*зоопарк, цирк* и под.). Для ассоциативных полей стимулов, называющих домашних животных (овца, лошадь, свинья и под.) появление реакций типа А1 не характерно, но их место занимают реакции типа А2 (*хлев, конюшня, сарай* и под.), реакциям Б1 и Г1 вместе (среда обитания и место, где животные демонстрируются или где их обычно можно видеть) можно поставить в соответствие реакции типа Б2–Г2 (*деревня, пастбище, луг, поле, ипподром*) на стимулы баран, лошадь, корова и под. Уже здесь открывается возможность обобщения представлений о связанных с животными референтными сферами (место жизни, обитания и место, где животных можно видеть, наблюдать). Однако реакции, указывающие на референтную сферу «место», обнаруживаются в ассоциативных полях почти каждого из стимулов, представленных конкретными существительными: кастрюля → *кухня, плита, стол*; бочка → *огород, пивная, склад*; цветок → *горшок, ваза, луг, в поле, рынок*; документ → *папка, в кармане, в столе, архив*; рисунок → *рамка, альбом, на стене, в рамке, в тексте*; доктор → *госпиталь, больница, профилакторий*; фамилия → *паспорт, анкета, бумага, список, дневник, кабинет*; гора → *Кавказ, Гималаи, в пустыне*; радуга → *небо, на небе, в небе*; удочка → *озеро, пруд, вода, в воде, в руке* и др.

Учет этих и подобных им типов реакций позволяет заключить, что привязка к месту, а также времени (хотя и с существенно меньшей степенью обязательности), то есть хронотоп, образует одну из главных характеристик фрагментов мира, которая отражается в структуре ассоциативных полей стимулов-существительных. «Хронотоп» в предлагаемой модели относится к категориям высшего уровня обобщений, однако и в данном, и в других подобных случаях во фрейме сохраняется вся иерархия признаков, так что модель может использоваться с разными уровнями генерализации.

Применение описанного выше подхода к систематизации ассоциативных реакций позволяет выделить и другие сферы референции, представленные в своем наиболее общем, категориальном, виде во всех рассмотренных ассоциативных полях. Одна из них — категория «Человек». Ее проявлени-

ями выступают, в частности, такие сферы: — «лицо, функционально связанное с предметом, который обозначен словом-стимулом» (ср.: лошадь → *наездник, всадник. ковбой, извозчик, казак, конюх* и др.; телефон → *абонент, подруга, ребенок* и под.; гора → *альпинист, скалолаз, турист, экстремалы* и под.; мел→ *учитель, школьник, маркёр*; — «принадлежность лицу» (ср.: лошадь → *моя, наша, ваша, друга, соседа, старика, Д'Артаньяна* и под.; телефон → *мой, мужа, друга* и под.; лоскут → *мой, себе*; роща → *наша*, ружьё → *моё* и под; — «использование человеком» (ср.: лошадь → *ехать, кататься, работа, ломовая, беговая, скачки* и под.; телефон → *звонить, позвонить, разговаривать, переговоры, общение, поболтать* и под.; гора → *отдых, чистый воздух, поход, покорение, кататься* и под.; лоскут→ *закладка, для платья, заплатка, игрушка, мыть пол, пришить, уборка*; мел → *грызть, для доски, красить, кушать, мелить, писать, рисовать*; — «артефакт, функционально ориентированный на предмет, который обозначен словом-стимулом» (ср.: лошадь → *телега, повозка, сани, седло, хомут, кнут, подкова* и др.; телефон → *провод, линия, модем, сеть, факс* и под.; гора → *рюкзак, лыжи, санки* и под.; лоскут→ *нитка, иголка, кукла, машинка швейная, мешок, одеяло*; мел → *доска, доска в школе, рисунок, тряпка*.

Другими проявлениями той же категории «Человек» являются сферы «действие человека на предмет, обозначенный словом-стимулом» и «действие предмета, обозначенного словом-стимулом на человека» (например: корова → *пасти* и корова → *бодает*, *бодается*).

Ассоциативные поля не всех стимулов-существительных содержат полный набор проявлений соответствующей категории. Так, обращаясь к приведенному выше ассоциативному материалу категории «Связь с человеком», можно отметить, что в корпусе ответов испытуемых на стимул лоскут не нашлось номинаций лиц, функционально связанных с предметом, который обозначен данным стимулом, среди рассмотренных реакций на стимулы гора и мел не нашлось реакций выражающих принадлежность соответствующих предметов лицам. Подобные факты требуют по крайней мере двух замечаний.

Первое: о представленности той или иной категории в ассоциативном поле конкретного стимула можно говорить, если реакциями на данный стимул фиксируется хотя бы одно из предусмотренных моделью частных проявлений рассматриваемой категории, а то, какие именно из частных проявлений категории представлены или не представлены в ассоциативном поле стимула, может использоваться как классификационный признак слов-стимулов и их содержания. При этом существенны не только качественные, но и количественные характеристики. Например, реакции относящиеся к категории «Качества, оценки предмета, обозначенного словом-стимулом», обнаруживаются во всех ассоциативных полях, но в полях одних стимулов чаще других встречаются эстетические оценки, в полях других стимулов — параметрические, в третьих — прагматические, эмоциональные, морально-этические, цветовые и другие.

Второе: если в достаточно представительном ассоциативном материале не зафиксированы реакции, относящиеся к тому или иному конкретному проявлению категории, это не означает, будто этот параметр вообще отсутствует в картине мира носителей языка. Он может быть убедительно документирован текстами на данном языке или обнаружен путем интроспекции. Подобные лакуны говорят не о слабости ассоциативного метода, как иногда утверждается исследователями тематической организации словаря, а, скорее, о его тонкости. В текстах тема, задаваемая тем или иным стимулом, может получать бесконечное развертывание в самых разных направлениях и аспектах, в том числе не только в общих, но и в частных, применительно не только к типичным ситуациям, но и к редким, случайным, даже исключительным. В отличие от текстовых материалов свободный ассоциативный эксперимент обычно представляет тему ограниченнее и одновременно актуальнее в психологическом, социальном и культурно-историческом плане.

Так, в материалах РАС и АССШС ассоциативные поля стимулов цветок и роза включают (помимо других реакций) ряд номинаций лиц. Это номинации тех, кому цветы дарят, и номинации самих дарителей, но не тех, кто цветы выращивает, за ними ухаживает, кто их изучает, выводит, кто их рисует, ими торгует и т. п. Есть реакции *рынок* и *рубль с полтиной*, то есть реакциями отмечена в ряду других и ситуация покупки цветов, но не названы продавцы, торговцы в качестве обязательных ее участников, как не названы и цветоводы, ботаники, агрономы, художники, аранжировщики букетов, декораторы, дизайнеры и другие лица, имеющие то или иное отношение к цветам, хотя такими номинациями язык обладает и нетрудно найти тексты с их использованием. Предмет, выделяемый стимулом, может быть участником множества ситуаций, и его место в них, роль, значимость, типичность оказываются неодинаковыми, неодинакова также актуальность самих различных ситуаций для участников эксперимента. Следствия этого мы и обнаруживаем в структуре ассоциативных полей, поэтому их структура не требует искусственного дополнения, выравнивания, а просто должна быть учтена. В рассмотренном выше случае аспект символического использования цветов оказывается для испытуемых актуальнее некоторых других.

На данном этапе работы основные направления ассоциативных связей исследуемых стимулов можно представить в следующей схеме (см. схему 1).

Схема 1. Сферы референции как концептуальные переменные образа мира[1]

1. Ситуации с участием S.
2. Предметно-логические связи S:
2.1. суперординаты разных уровней;
2.2. субординаты (разновидности S);
2.3. предметы того же рода, что и S;
2.4 противоположности, антиподы S;
3. Меронимы S.
4. Хронотоп:
4.1. типичное место;
4.2. типичное время.
5. Аксессуары S (сопутствующие S предметы в типичных ситуациях).
6. Качества, оценки S.
6.1. параметрические;
6.2. по материалу, форме, цвету;
6.3. прагматические;
6.4. эстетические;
6.5. эмоциональные
и другие.
7. Действия, состояния, деятельность.
8. Связь S с человеком:
8.1. лица, функционально связанные с S;
8.2. принадлежность лицу;
8.3. использование S человеком;
8.4. артефакт, функционально ориентированный на S;
8.5. действие человека на S;
8.6. действие S на человека.
9. Речевой аспект S:
9.1. индивидуальное (собственное) имя S;
9.2. синонимы слова-стимула;
9.3. метаязыковая характеристика слова-стимула;
9.4. метафорическое использование слова-стимула;
9.5. формулы, прецедентные тексты со словом-стимулом или с S;
9.6. слова, словоформы, близкие по звучанию слову-стимулу;
9.7. наука об S;
9.8. толкование стимула.
10. Символическое использование S.

В качестве примера соотнесения ассоциативных реакций со слотами указанного фрейма приведем материал ассоциативного поля ПАУК:

1. *я увидел паука, там паук*… 2.1. *животное, насекомое, букашка*… 2.2. *тарантул, птицеед*… 2.3. *таракан, клещ, клоп*… 2.4. *муха*; 3. *восемь ног, жало*… 4.1. *угол, в доме, на стене*… 5. *паутина, грязь*… 6.1. *большой, маленький*… 6.2. *серый, черный*… 6.3. *ядовитый,* 6.4. *красивый, безобразный*… 6.5. *гадость, противный*… 7. *затянул, ползет, затаился, ест*… 8.2. *мой, домашний*… 8.4. *тапок*; 8.5. *убить*, 8.6. *боязнь, боюсь, кусает*… 9.2. *мизгирь,* 9.4. *человек, злодей, мафия*… 9.5. *Муха-цокотуха, Шнюк, ой!, ай!, фу!*... 9.6. *ук, лук, наук*… 9.7. *биология*, 9.8. *насекомый паукообразный*, 10.*плохая примета, знак, весть, к письму*…

Необходимо специально подчеркнуть, что задача приведенной схемы не классифицировать предметы, понятия или слова (в этом ее принципиальное отличие от структурной основы идеографических и семантических словарей), а показать, с каких точек зрения рассматриваются людьми предметы и из каких образов составляются соответствующие картины мира. Полученная структура сопоставима с ассоциативными гештальтами Ю. Н. Караулова [Караулов 2002], но не совпадает с ними по содержанию и — главное — не строится на основаниях, привнесенных из системы местоимений, а использует принципы, устанавливаемые на материале самих ассоциаций.

4. Заключение

Категории, представленные в схеме, настолько тесно связаны между собой, что не только не исключают одна другую, но проявляют свойства взаимообусловленности и для части стимулов могут совмещаться, так что их разграничение, хотя оно и необходимо, является в определенной мере условным. Так, категория «Связь предмета, обозначенного словом-стимулом, с человеком» рассматривается как самостоятельная, поскольку лица представлены в ассоциативных реакциях в явном виде, выступают участниками типичных ситуаций с предметом, человек является творцом артефактов, ориентированных на предмет, и т. д. Однако и качества предметов выделяются людьми, и оценки, и предметно-логические квалификации, и сама речь принадлежат человеку. Человеческий аспект проявляется и в принципиальной неполноте, избирательности образов, которые репрезентируются ассоциативными реакциями и отражают специфику картины мира в человеческом сознании.

Ассоциативные реакции могут называть целые ситуации (цветок → *свидание, День рождения*; мел → *урок*; фамилия → *знакомство*, *замужество*, *проверка*; доброволец → *фронт*, живот → *беременность*; и под.), но гораздо чаще ассоциации отмечают их компоненты, стороны, обстоятельства, представленные в схеме отдельными категориями. С этой точки зрения предложенная схема непоследовательна: целое и части выступают в ней на одном и том же уровне обобщения, но это особенность самих ассоциативных структур, и по крайней мере на данном этапе работы, по-видимому, нет оснований от неё избавляться.

[1] В данной схеме символ «S» имеет значение: ‘предмет, обозначенный словом-стимулом’.

Исследование, проведенное на материале русских ассоциативных словарей, показало, что на самом деле существует универсальный комплекс концептуальных переменных, организующих образы предметов в русском коллективном сознании, и необходимо продолжить изучение данного явления на материале ассоциативных полей стимулов других семантических типов.

Литература

1. *АСШС 2009 — Гольдин В. Е., Мартьянов А. О., Сдобнова А. П.* Электронный русский ассоциативный словарь школьников // Компьютерная лингвистика и интеллектуальные технологии (по материалам ежегодной Международной конференции «Диалог 2009»). Вып. 8 (15). М.: 2009.
2. *Гасица Н. А.* Ассоциативная структура значения слова в онтогенезе: Автореф. дис. на соиск. учен. степ. канд. филол. наук: 10.02.19. М., 1990.
3. *Караулов Ю. Н.* Национальные образы сознания в ассоциативной структуре слова // Национальный менталитет и языковая личность. Пермь: 2002.
4. *РАС 2002* — Русский ассоциативный словарь. М.: АСТ, Астрель. Т. 1–2.

Вокальный жест *А* в устной речи[1]

Vocal Gesture *Ah* in Spoken Russian

Гришина Е. А. (rudi2007@yandex.ru)

Институт русского языка РАН (Москва)

В статье на материале Мультимедийного русского корпуса (МУРКО) анализируется употребление вокального жеста *А*. Для анализа привлекаются сведения о типе жестикуляции, которая сопровождает этот вокальный жест в речи. В результате исследования у вокального жеста *А* обнаруживается четыре типа употреблений — вопросительное (*А* в значении вопросительной частицы), в значении пренебрежения (А в значении отрицательной частицы), ментальное (устное междометие *А* в значении понимания) и физиологическое (*А* как возглас). В статье также производится сопоставление вокальных жестов *А* и *О*.

[Марстон] *Мистер Дэвис!*
[Блор] *А?*
[Марстон] *А!*
[Блор] *А… Да! Пожалуй / дальше играть в прятки нечего.*
Десять негритят, С. Говорухин, 1987

[Степан] *Начальство.*
[Серафим] *А?*
[Степан] *Хозяйка!*
[Серафим] *Ааа!*
[Степан] *Ууу!*
[Серафим] *Ооо! Пошли.*
Стряпуха, Э. Кеосаян, 1965

1. Постановка задачи

В работе [Гришина 2009] на материале Мультимедийного русского корпуса (МУРКО) нами было проанализировано употребление вокального жеста *О* в русском устном диалоге. Для этого была предложена специальная методика, которая предполагает анализ употребления вокального жеста на основе тех телесных жестов (телесными жестами, в отличие от вокальных жестов, мы называем обычные мимику и жестикуляцию), которые его сопровождают. В настоящей работе мы хотим применить эту же методику к анализу вокального жеста *А*, а также сопоставить полученные результаты с результатами, касающимися вокального жеста *О*.

Напомним, что при анализе вокального жеста *О* были приняты следующие допущения. Во-первых, предполагалось, что в том случае, если то или иное языковое явление постоянно сопровождается некоторым классом однотипных жестов, связь между этим языковым явлением и этим типом жестов не является случайной — а именно, значение телесных жестов в этом случае совпадает с некоторой зоной значения данного языкового явления. Следовательно, учет жестикуляции следует рассматривать как необходимый компонент при анализе любого аспекта устной речи.

Во-вторых, при анализе устного диалога жесты не должны рассматриваться поодиночке, поскольку для повседневной жестикуляции характерен высокий уровень синонимии между жестами самой разной структуры и места образования: жесты должны организовываться в кластеры. В кластерах следует выделять ***базовый жест***, значение которого входит в качестве семантического примитива в состав значений ***производных жестов*** (так, например, значение жестов, выражающих удивление говорящего, является базовым по отношению к жестам, выражающим высокую оценку или сопровождающим приветствие).

В настоящей работе мы будем также применять понятие ***доминантного жеста*** — это жест, кото-

[1] Исследование проведено при поддержке грантов РФФИ 08-06-00371а, 10-06-00151а, а также программы ОИФН РАН 2009–2011 гг. «Генезис и взаимодействие социальных, культурных и языковых общностей».

рый чаще всего сопровождает то или иное языковое явление (в нашем случае — тот или иной вокальный жест). Например, для выражения согласия очевидным доминантным жестом является кивок. Доминантных жестов может быть больше одного.

2. Анализ материала

Вокальный жест *А* достаточно полно описан в научной литературе и в словарях. Так, в работе [Шаронов 2006] для него предложено три основных значения:

1) Реакция на физическое воздействие или непосредственную опасность
2) Реакция на окончательное осознание чего-л.
3) Реакция на осознание незначительности чего-л.

Это частично совпадает с описанием, предложенным в Словаре Ушакова:

А[4], *междом.* **1.** Выражение догадки, удивления (произносится протяжно) (≈ пункту 2 в классификации И. А. Шаронова); **2.** Выражение решимости с оттенком отчаяния или досады (≈ пункту 3 в классификации И. А. Шаронова).

Кроме того, в Словаре Ушакова выделено еще два значения *А*:

А[4], *междом.* **3.** При повторном обращении к кому-н., как выражение нек-рой настойчивости. *Ваня, а Ваня, иди сюда!*,

а также

А[3], *вопросит. частица*. Что? как? *Что ты сказал, а?* [Ушаков 1935].

В Словаре [Ожегов, Шведова 1999] информация, зафиксированная в Словаре Ушакова, несколько переформатирована, в частности, третье значение *А*[4] в Словаре Ушакова (повторное обращение) считается такой же частицей, как и *А*[3] в Словаре Ушакова (и это, с нашей точки зрения, верно), но при этом такой же частицей считается второе значение в классификации И. А. Шаронова и первое значение *А*[4] в Словаре Ушакова, т. е. ‘*А* понимания’ (см. ниже) выводится из состава междометий. Кроме того, в [Ожегов, Шведова 1999] не фиксируется *А* в значении решимости-отчаяния-досады-осознания незначительности (пункт 3 в классификации И. А. Шаронова и второе значение *А*[3] в Словаре Ушакова), однако вводится *А*, которое, судя по приведенным контекстам, близко к первому пункту в классификации И. А. Шаронова (и только это *А* признается в данном словаре междометием).

В совокупности, если отвлечься от разнобоя в частеречной атрибуции и в дефинициях, вообще характерных для словарного описания междометий, состав значений вокального жеста *А* перечислен достаточно полно, что и подтверждает наш материал.

2.1. Основные типы телесных маркеров, сопровождающих вокальный жест *А*[2]

2.1.1. Группа ‘вопрос’.

Базовыми жестами в этой группе употреблений вокального жеста *А* являются телесные жесты со значением ‘установление контакта’ (*перевести взгляд на адресата, обвести взглядом собеседников, двинуть головой вперед*[3]*, коснуться собеседника*). При этом распределены эти жесты достаточно четко: в случае, если перед употреблением вокального жеста *А* говорящий уже смотрел на собеседника, то он использует жесты *двинуть головой вперед, коснуться собеседника,* т. е. жесты, актуализирующие уже существующий контакт; если же он смотрел в сторону, то используется жест *перевести взгляд на собеседника*[4]; если адресатов высказывания несколько, то может использоваться жест *обвести взглядом собеседников*.

Доминантными телесными жестами в этой группе употребления вокального жеста *А* являются (по мере убывания частотности) *перевести взгляд, поднять брови, двинуть головой вперед.*

На основе этого базового значения формируются следующие производные значения:

1) ***‘отклик’***: в ответ на обращение говорящий с помощью вокального жеста *А* и сопровождающих жестов демонстрирует, что он *установил контакт* с обращающимся и готов к дальнейшему общению.

(49) [Сан Саныч] *Степа!*

Речевой ряд	[Степан] ***А?***
Событийно-жестовый ряд	переводит взгляд на Сан Саныча

[Сан Саныч] *Ты лучше за мотором следи!*

Спортлото-82, Л. Гайдай, 1982

(50) [Петр] *Няня Зоя! Няня Зоя!*

Речевой ряд	[Зоя] ***А?***
Событийно-жестовый ряд	наклоняется к Петру

[Петр] *В этом месяце тридцать дней?*

Актриса, Л. Трауберг, 1942

---

[2] Отметим, что в настоящей статье из рассмотрения выводятся случаи употребления вокального жеста *А* на вдохе, а не на выдохе (аргументацию мы приводили при рассмотрении вокального жеста *О*).

[3] Трактовка жеста *двинуть головой вперед* как жеста ‘установления контакта’ требует некоторого пояснения. Как будет указано позже (см. раздел 2.1.5, п. 2), осуществляя этот жест, говорящий придвигает голову ближе к адресату, чтобы как бы убыстрить получение искомой информации. Тем самым говорящий, приближаясь к адресату, привлекает к себе его внимание и стимулирует более оперативную реакцию адресата.

[4] Если за *А* следует или *А* предшествует обращение, то жест *перевести взгляд* совпадает с обращением, а не с вокальным жестом *А*.

(3)

Речевой ряд	[Эпштейн] *Ножницы.*	[Куркова] ***А?***	*Угу.*
Событийно-жестовый ряд	подсказывает Курковой, показывает пальцем	переводит взгляд на Эпштейна	понимает, берет ножницы

Операция «С Новым Годом!», А. Рогожкин, 1996

(4) [Верочка] *Вы что / репетировали?* [Лисичкин] *Нет / дочь моя / я не репетировал.*

Речевой ряд	[Верочка] *А!*	[Лисичкин] ***А?***
Событийно-жестовый ряд	видит, что Лисичкин загримирован, пугается, отворачивается, зажмуривается	не понимает такой реакции Верочки

Ах, водевиль!, Г. Юнгвальд-Хилькевич, 1979

(5) [Алексей Борисович] *А у тебя нет такого приспособления / чтоб всех троих нас снять?*

Речевой ряд	***А?***	[Владимир Сергеевич] *Нет / такой штуки нет.* [Алексей Борисович] *Слышишь! Всех троих! Вместе с псом!*	***А?***
Событийно-жестовый ряд	говорящий не попадает в кадр		двигает головой вперед

Послесловие, М. Хуциев, 1983

2) ***‘не слышу’***: в ответ на некоторый речевой стимул говорящий с помощью вокального жеста *А* и сопровождающих жестов демонстрирует автору стимула, что он не расслышал или недопонял его реплику, и, *устанавливая контакт* с автором стимула, просит повторить или прояснить стимул [см. (3)].

Довольно часто в этом случае в качестве недопонятого стимула выступает некоторое событие или действие адресата [см. (4)].

3) ***‘побуждение к ответу’***: говорящий, обращаясь к слушающему, побуждает его более оперативно реагировать на стимул, и для этого актуализирует уже *установленный* к данному моменту *контакт* со слушающим [см. (5)].

В данном значении вокальный жест *А* входит в группу дискурсивных маркеров, или маркеров разговорной речи, которые характерны для русской устной речи и, по нашим данным (см. [Гришина 2007]), появляются при переходе от письменных текстов к устным (например, при переходе от сценария/пьесы к фильму) и исчезают при движении в обратном направлении (например, при переходе от фильма к субтитрам). В позиции в конце фразы вокальный жест *А* используется наряду с маркерами *ну* и *да*, например:

(6) [Медсестра Куркова] *Ну зачем вы встали? Лежали бы!* ***Ну?***
[Безнадежно больной] *Новый Год…*

Операция «С Новым Годом», А. Рогожкин, 1996

(7) [Афоня] *Ты из АТС / с набережной?*
[Коля] *Нет.*
[Афоня] *Из третьего СМУ? Вася?* ***Да?***
[Коля] *Нет. Я штукатур. Коля.*

Афоня, Г. Данелия, 1975

Все эти три маркера достаточно часто инкорпорируются в состав предшествующего стимула, так что межфразовая пауза между стимулом и маркером заменяется межсловной:

(8) [Иванова] *Интересно всё /* ***да?*** *В новинку /* ***а?***

Друг мой, Колька!, А. Салтыков, 1961

(9) [Пионерка] *Ну кто ж за тебя будет бином Ньютона учить /* ***а?***
[Пимен] *Опять за свое! Бином-бином / бином / ну скока можно /* ***ну?***
[Пионерка] *А ну пошли в класс! А ну пошли в класс!*
[Пимен] *Не пойду я /* ***ну!***

Друг мой, Колька!, А. Салтыков, 1961

Эти маркеры устной речи выполняют сходные функции — побуждают слушающего к скорейшей реакции на стимул и демонстрируют заинтересованность говорящего в реакции слушающего, — но имеют различное происхождение.

Основное значение ударного *ну* в устной речи — побуждение слушающего к действию или речи, таким образом, в конце фразы ударное *ну* вынуждает слушающего интенсивней реагировать на эту фразу.

Ударное *да* в этой позиции предвосхищает положительную реакцию слушающего на фразу-стимул, как бы подсказывает слушающему желательную для говорящего положительную реакцию и тем самым пытается ускорить ее.

Ударное же *а* в этой конструкции восходит к вокальному жесту *А* со значением ‘не слышу’ (см. выше, п. 2): в ситуации, когда говорящий задает слушающему некоторый вопрос (делает утверждение, побуждает слушающего к совершению какого-

либо действия) и реакция от слушающего поступает с задержкой (конечно, с точки зрения говорящего), говорящий как бы считает, что реакция от слушающего поступила, но он, говорящий, ее не расслышал — и соответственно, с помощью *а* он побуждает слушающего повторить ответ. Понятно, что такая схема является удобным способом подтолкнуть слушающего к оперативной реакции на реплику говорящего, даже если никакой объективной задержки реакции со стороны слушающего не было.

Объяснения требует тот факт, что все три группы значений вокального жеста *А* ('отклик', 'не слышу', 'побуждение к ответу'), реализуются в основном в форме вопроса. По-видимому, здесь мы сталкиваемся с широко распространенным в русской устной речи процессом свертывания конструкций. В работах [Гришина 2007], [Grishina 2007] мы уже писали о том, что метатекстовые вставки в устном нарративе, которые включают частицу *вот*, являются свернутыми до начальных элементов метатекстовыми фразами. Так, например, открывающие метатекстовые вставки *Во'т что, Так во'т* являются результатом свертывания фразы (*Так*) *Вот что я тебе сейчас скажу*; закрывающие метатекстовые вставки *Вот та'к, Та'к вот* являются результатом свертывания фраз *Вот так обстоят дела, Так вот обстоят дела*. В работе [Гришина 2008] было показано, что и сама частица *вот* в изолированном употреблении (например, между межфразовыми паузами, или в абсолютном начале текста перед межфразовой паузой) представляет собой подведение итога для всего сказанного непосредственно перед *вот*, т. е. является не первичным дейксисом, а свернутым высказыванием[5].

Мы предполагаем, что в группах 'отклик' и 'не слышу' за вокальным жестом *А* следует свернутый до нуля, вернее, до вопросительной интонации вопрос:

для 'отклик'	≈Что тебе нужно?
для 'не слышу',	≈Что ты сказал?,
'не понимаю'	≈Что случилось?

Таким образом, в развернутом виде использование вокального жеста *А* в этих контекстах выглядит следующим образом (свернутый элемент дан нижним индексом):

(1а) [Сан Саныч] *Степа!*
[Степан] *А* $_{\text{Что тебе нужно}}$*?*
(2а) [Эпштейн] *Ножницы*.
[Куркова] *А* $_{\text{Что ты сказал}}$*?*
(3а) [Верочка] [пугается] *А!*
[Лисичкин] *А* $_{\text{Что случилось}}$*?*

Именно такая структура подразумеваемого вопроса, от которого осталась только вопросительная интонация, объясняет, почему в контекстах этого типа вокальный жест *А* легко и без потери смысла может быть заменен на вопросительное *что?*:

(1б) [Сан Саныч] *Степа!*
[Степан] *Что?* (=*Что* $_{\text{тебе нужно}}$*?*)
(2б) [Эпштейн] *Ножницы*.
[Куркова] *Что?* (=*Что* $_{\text{ты сказал}}$*?*)
(3б) [Верочка] [пугается] *А!*
[Лисичкин] *Что?* (=*Что* $_{\text{случилось}}$*?*)

Что касается группы 3, 'побуждение к ответу', то здесь вопросительная интонация задается филиацией значения (а именно, тем фактом, что 'побуждение к ответу' восходит к значению 'не слышу'), а невозможность замены *а?* на *что?* в данной конструкции (**Ну кто ж за тебя будет бином Ньютона учить / **что?***) объясняется, по-видимому, тем, что данное употребление вокального жеста *А* является своего рода коммуникативным фразеологизмом, а как известно, замена элемента фразеологизма на синонимичный часто разрушает фразеологизм (*закинуть удочку* vs. **закинуть спиннинг*).

Именно вопросительная интонация всех перечисленных групп ('отклик', 'не слышу', 'побуждение к ответу') позволила нам объединить их в один тип — 'вопрос'.

2.1.2. Группа 'понимание'.

1) Базовыми жестами в этой группе являются жесты, которые сопровождают процесс получения и усвоения человеком некоторой информации, т. е. ***'понимание'*** некоторого факта (*откинуть голову, двинуть головой назад, поднять брови, кивнуть, вздрогнуть, вскинуть руки, поднять палец, коснуться головы* и некоторые другие). Доминантными жестами являются (по мере убывания частотности) *откинуть голову, поднять брови, кивнуть*.

(10) [Инга] *Привет*. [Гоголев] *А я щас был у твоей квартирной хозяйки.*

Речевой ряд	[Инга] *А /*	*ты с ней познакомился?*
Событийно-жестовый ряд	откидывает голову назад	

[Гоголев] *Познакомился*.

Американский дедушка, И. Щеголев, 1993

(11) [Крошкин] *Можно мне повидать ефрейтора Калмыкову?* [Тихон] *Их нет / товарищ старший лейтенант.*

[5] Этот факт был проницательно отмечен в работе [Гольдин 1998:42], однако автор объясняет его тем, что в значение *вот* входит в качестве необходимого компонента анафоричность. С нашей точки зрения, в приписывании *вот* имманентной анафоричности нет необходимости — поскольку изолированное *вот* является свернутым высказыванием, то и анафоричность его значения может рассматриваться как производная. Именно «пустота» *вот* как дейксиса, в том числе и отсутствие в нем анафоричности, объясняет универсальность использования этой указательной частицы в устной речи.

Речевой ряд	[Крошкин] *А /*	*ну ничё / работайте / работайте.*
Событийно-жестовый ряд	кивает	

Беспокойное хозяйство, М. Жаров, 1946

2) Производными жестами являются жесты ***‘интеллектуального удовлетворения’***, когда говорящий испытывает удовлетворение от *понимания* некоторого факта, от получения той или иной информации:

(12) [Эмили Брент] *Я знаю / как зовут убийцу. Ее зовут… Беатрис Тейлор.*

Речевой ряд	*А!*	
Событийно-жестовый ряд	откидывает голову назад	смеется, кивает, поднимает палец

Десять негритят, С. Говорухин, 1987

3) Также производными от базовых жестов со значением ‘понимания’ являются жесты ***‘узнавания’***, сопровождающие осознание, *понимание* говорящим появления в его поле зрения, зоне внимания некоторого нового объекта или человека:

(13) [Вера] [у закрытой двери] *Кто? Сирил / ты?*

Речевой ряд	*А /*	*Сирил! Сейчас / малыш!*
Событийно-жестовый ряд	откидывает голову	

Десять негритят, С. Говорухин, 1987

2.1.3. Группа ‘пренебрежение’.

1) Следующей крупной группой употреблений вокального жеста *А* является группа, основанная на кластере жестов, базовым значением который является ***‘пренебрежение’*** (*махнуть рукой, сморщиться, отвернуться, отбросить что-л., пожать плечами, поднять брови* и некоторые другие). Используя эти жесты, говорящий описывает какую-то ситуацию как не стоящую внимания, не заслуживающую серьезного рассмотрения. Доминантным жестом в этой группе является жест *махнуть рукой* (*отмахнуться*).

(14) [Бармалей] *Доктор*! [Разбойник] *Доктор!*

Речевой ряд	[Айболит] *Да-да?*	*А!*
Событийно-жестовый ряд	оглядывается, видит разбойника	отмахивается

(15)

Речевой ряд		[Блор] *Что такое?*	*А!*	*Мотор не работает.*
Событийно-жестовый ряд	Блор пытается включить свет, свет не включается	оглядывает комнату	машет рукой	

Десять негритят, С. Говорухин, 1987

2) Производными жестами являются жесты ***‘досады’***, когда говорящий с *пренебрежением* отказывается от чего-то настолько плохого, что от него невозможно ждать никаких положительных перспектив и возможностей [см. (15)].

3) Также производными являются жесты ***‘решительности’***, когда говорящий принимает решение предпринять какое-то действие, с *пренебрежением* отнесшись к грозящей ему опасности:

(16) [Сан Саныч] *Много текста! Иди ищи!* [Степан] *Да как я най…* [Сан Саныч] *Иди!*

Речевой ряд	*А /*	*чёрт!*
Событийно-жестовый ряд	чешет затылок, морщится	уходит

Спортлото-82, Л. Гайдай, 1982

2.1.4. Физиологическая группа.

Как и в случае вокального жеста *О,* группа жестов, сопровождающих вокальный жест *А* в физиологических значениях, не является однородной и подразделяется на несколько подгрупп, в разной степени связанных между собой.

1) Подгруппа ***‘ужас’***. Сопровождающие жесты — *стараться убежать, подскочить, расширить глаза, закрыть лицо руками, прикрыть рот рукой, отшатнуться, отбросить что-л.*:

(17)

Речевой ряд	[Каин XVIII] *А! А! А! А! Ай!*	*Как мне страшно!*
Событийно-жестовый ряд	расширяет глаза, пытается убежать	

[Туалетный работник] *Что с вами / ваше величество?*

Каин XVIII, Н. Кошеверова, 1963

2) Подгруппа ***‘неожиданность’***. Сопровождающие жесты — *вздрогнуть, схватить собеседника за руки, отшатнуться, подпрыгнуть, отвернуться, прижать руку ко рту*:

(18) [Наташа будит Кузьму]

Речевой ряд	[Кузьма] *А! А! А!*
Событийно-жестовый ряд	резко подскакивает

[Наташа] *Вредно спать так.*

Когда деревья были большими, Л. Кулиджанов, 1961

(19)

Речевой ряд	[Качалов] *Э! Ё… Всё!*	[Мавецкий] *А!*	*А!*	*Мать вашу…*
Событийно-жестовый ряд	вправляет Мавецкому челюсть	говорящий не попадает в кадр	держится за челюсть	

Операция «С Новым Годом!», А. Рогожкин, 1996

Как видим, группы 1 и 2 в отношении жестов имеют много общих элементов. Но и в речевом отношении, с точки зрения употребления вокального жеста *А*, эти группы пересекаются: реакция на неожиданность часто сопровождается испугом, а разница между испугом и ужасом не принципиальна: ужас, во-первых, интенсивней испуга, а во-вторых испуг чаще связан с неизвестностью (говорящий пугается чего-то неожиданного и неизвестного), а ужас связан с осознанием говорящим того, что с ним происходит. Как станет ясно далее (см. раздел 2.3.3), разделение этих двух физиологических реакций на две близкие, но разные группы обоснованно — если не различием и плотностью сопровождающих жестов, то интонационными и фонетическими особенностями каждой группы.

Основными жестами в двух этих подгруппах являются жесты, связанные с желанием удалиться, отделиться от чего-то (*бежать, подскочить, отшатнуться, отвернуться, зажмуриться, закрыть лицо руками, отбросить что-л.*), а также с желанием сделаться незаметным (*прижать руки ко рту, присесть*). Ввиду небольшого количества примеров выделить доминантные жесты не удается.

3) Подгруппа ***‘боль’***. Сопровождающие жесты — *схватиться за больное место, сморщиться, подскочить* [см. (19)].

Доминантным жестом в этой группе является жест *держаться/схватиться за больное место*.

4) Подгруппа ***‘интенсивность чувств’***. Сопровождающие жесты — *всплеснуть руками, бросить что-л., расширить глаза, щелкнуть собеседника по лбу:*

(20) [Крошкин] *В чем дело / Жан?*

Речевой ряд	[Лярошель] *А!*
Событийно-жестовы ряд	бросает парашют

[Крошкин] *Почему вы решили плюхнуться здесь?*
[Лярошель] *Я немного увлекался на бой. Кончилось горючий.*

Беспокойное хозяйство, М. Жаров, 1946

Ввиду недостаточного количества примеров выделить доминантный жест не удается.

Таковы основные типы употребления вокального жеста *А*, если использовать для их вычленения сопровождающие вокальный жест *А* телесные жесты.

2.1.5. Доминантные жесты: комментарии.

Как видим, три первые группы употреблений вокального жеста *А* имеют свои доминантные жесты (четвертую, физиологическую группу, в таблицу не включаем, поскольку недостаточное количество примеров не дает возможности надежно выделить доминантный жест)[6]:

Таблица 1

Жесты	**группа ‘вопрос’**	**группа ‘понимание’**	**группа ‘пренебрежение’**
перевести взгляд	+	–	–
поднять брови	+	+	–
двинуть головой вперед	+	–	–
откинуть голову	–	+	–
кивнуть	–	+	–
махнуть рукой (отмахнуться)	–	–	+

Некоторые комментарии к таблице.

1) Как видим, основным контактоустанавливающим жестом является *перевести взгляд*, т. е. зрительный контакт преобладает над слуховым, невзирая на то, что устная речь — явление звуковое (т. е. логичнее было бы, если бы основным контактоустанавливающим жестом здесь был жест *придвинуть ухо к собеседнику*, который, действительно, используется как сопровождение вокального жеста *А* в этом значении, но настолько редко, что его употребление может считаться маркированным, например, стилистически).

2) Интересна антонимия жестов *двинуть головой вперед* (для группы ‘вопрос’) и *откинуть голову назад* (для группы ‘понимание’). Если этимология употребления первого жеста, *двинуть головой вперед*, достаточно очевидна, — говорящий перемещает голову ближе к собеседнику, чтобы требуемая информация как бы быстрее попадала в его мозг, то жест *откинуть голову назад* как знак понимания не имеет, с нашей точки зрения, очевидного объяснения. В качестве версии можно предложить обратное движение головы от собеседника — в случае, если запрошенная ранее информация получена и понята (аналогом на физиологическом уровне является жест *отодвинуться от стола, откинуться на спинку стула* при насыщении).

[6] В таблице заштрихованы серым ячейки с самым частотным жестом в данной группе.

3) Жест *махнуть рукой* как знак пренебрежения (обычно сверху вниз или немного в сторону, иногда вперед, в направлении собеседника) имеет достаточно очевидную этимологию — он восходит к отбрасыванию чего-либо ненужного, неприятного и под.

4) Комментариев требует тот факт, что жест *поднять брови* входит в доминантную группу и для значения ‘вопрос’, и для значения ‘понимания’, т. е. в две группы, которые по другим доминантным жестам (см. п. 2) являются антонимичными. Объяснить это можно тем, что жест *поднять брови* является доминантным при выражении удивления (в частности, анализ показал, что именно этот жест является самым частотным среди телесных жестов, сопровождающих вокальный жест *О* в междометном значении, т. е. в значении удивления[7]). Поскольку и группа ‘вопрос’, и группа ‘понимание’ содержат в себе элемент удивления (говорящий удивляется чему-то и задает вопрос, получив же ответ и поняв его, он может испытать удивление от полученной информации), использование одновременно с вокальным жестом *А* в значении ‘вопрос’ и в значении ‘понимание’ одного из самых частотных жестов удивления совершенно логично.

[7] Связь жеста *поднять брови* с удивлением определяется тем, что этот жест — производный от жеста *расширить глаза*, более того, *поднять брови* — жест, до которого редуцируется жест *расширить глаза* при неполном осуществлении (т. е. нельзя расширить глаза, не подняв при этом брови, но можно поднять брови, не расширяя глаз, так что при редуцированном осуществлении жест *поднять брови* как бы представляет жест *расширить глаза*). Употребление же в состоянии удивления жеста *расширить глаза* достаточно понятно — жестикулирующий старается лучше рассмотреть и понять удивившее его явление.

2.2. Фонетические и интонационные особенности разных типов употребления вокальных жестов *А* и *О*

2.3.1. Анализ материала с помощью программы Speech Analyzer[8], которая, помимо прочего, наглядно показывает движение тона на гласных, дал для вокальных жестов *А* и *О* результаты, отраженные в Таблице 2.

Для того чтобы дать содержательную интерпретацию полученным данным, необходимо провести некоторые разграничения.

[8] © SIL International, Даллас, Техас, США.

Таблица 2

Вокальный жест	Значение	Нисходящий тон*	Восходяще-нисходящий тон**	Восходяще-нисходящий тон с падением***	Волновое движение тона****	Восходящий тон*****	Ровное движение тона******	Прочее
А и *О*	возглас*******	**34 %**	*11 %*	37 %	48 %	*5 %*	10 %	2 %
О	удивление	*4 %*	**41 %**	36 %	**77 %**	*3 %*	**16 %**	0 %
	указание	**38 %**	*10 %*	*14 %*	*24 %*	**36 %**	*0 %*	2 %
А	понимание	19 %	26 %	**45 %**	**71 %**	*7 %*	*0 %*	3 %
	вопрос	4 %	2 %	4 %	6 %	**88 %**	*0 %*	2 %
	пренебрежение	**55 %**	*10 %*	35 %	45 %	*0 %*	*0 %*	0 %
Всего		17 %	23 %	28 %	51 %	22 %	7 %	1 %

Пояснения к таблице:

1) полужирным шрифтом в заштрихованных ячейках отмечены данные, существенно отклоняющиеся от средних (т. е. от процентов в строке «Всего» в соответствующем столбце) **в бóльшую сторону**;

5) курсивом отмечены данные, существенно отклоняющиеся от средних *в меньшую сторону*.

* Нисходящим тоном считается движение тона по прямой, т. е. не включающее волновое движение (см. сноску ****), когда конечная точка находится заметно ниже начальной.

** Сочетание прямого (т. е. без волнового движения) восходящего тона (см. сноску *****) и нисходящего тона (см. сноску *).

*** Восходяще-нисходящим тоном с падением считается движение тона, при котором конечная точка спуска находится заметно ниже начальной точки подъема.

**** Объединяет те случаи, когда в процессе произнесения происходит возвратное движение тона, т. е. подъем на определенную высоту с последующим падением в исходную точку. Данный столбец является суммой двух предшествующих.

***** Восходящим движением считается движение, в котором конечная точка находится заметно выше начальной точки, вне зависимости от того, наличествует ли изменение тона (волновое движение) между начальной и конечной точками

****** Ровным движением тона считается движение тона без заметных изменений высоты.

******* Напомним, что в работе [Гришина 2009] возгласом считались употребления вокального жеста *О* в физиологических значениях (‘интенсивность чувств’, ‘боль’, ‘физическое напряжение’ и прочее). Так же трактуется и вокальный жест *А* в физиологическом значении (‘боль’, ‘неожиданность’, ‘ужас’, ‘интенсивность чувств’). О возгласах см. также ниже, разделы 2.3.2–2.3.3.

2.3.2. Выше (а также при анализе вокального жеста *О* в работе [Гришина 2009]) мы рассматривали разные случаи употребления вокальных жестов в устной речи как равноправные и иерархически соположенные. Представляется, что такая трактовка этих единиц не вполне правомерна.

1. Физиологические группы вокальных жестов *А* и *О* отличаются от остальных групп (*О* в значении 'удивления' и 'указания', *А* в значении 'понимания', 'вопроса' и 'пренебрежения') по трем параметрам:

1.1. Фонематическая атрибуция физиологических вокальных жестов не является достаточно четкой — есть некоторое количество контекстов, где трудно однозначно определить, с чем именно мы имеем дело, с *А* или с *О*, с *О* или с *У*.

1.2. Вокальные жесты *А* и *О* в физиологическом значении не являются знаками — без соответствующего контекста не удается определить, в какой именно ситуации исполняется этот вокальный жест: в ситуации, когда говорящий пытается передать интенсивность некоторого чувства или ощущения, испуг, боль или ужас. Соответственно, мы не можем заменить эти вокальные жесты языковыми синонимами и вынуждены для их описания использовать поясняющий метатекст (типа *закричал от боли, воскликнул от избытка чувств, завопил в ужасе* и под.).

1.3. И, наконец, вокальные жесты *О* и *А* в физиологическом значении сопровождаются телесными жестами низкой степени условности, т. е. максимально приближенными к утилитарным физиологическим движениям человека. Так, например, телесный жест *сморщиться* (*скривиться*) в случае вокального жеста *А* в значении 'пренебрежения' является условным жестом — он передает отвращение, испытываемое говорящим к некоторому явлению, ***как будто бы*** это явление было материальным и имело физиологически неприятный вкус или доставляло говорящему физиологически неприятное ощущение. Этот же телесный жест, сопровождающий вокальные жесты *А* и *О* в значении 'боль', используется говорящим в прямом, а не переносном значении, — именно как прямая реакция на болевые ощущения.

2. Вокальные жесты *А* и *О* в базовых значениях 'понимание' и 'удивление' по этим же параметрам характеризуются следующим образом:

2.1. В отличие от физиологических вокальных жестов, фонематическая природа этих вокальных жестов достаточно определенна — вокальный жест *О* в значении 'удивления' нельзя спутать ни с жестом *А*, ни с жестом *У*, а вокальный жест *А* в значении 'понимания' — с вокальным жестом *О*.

2.2. Как и физиологические вокальные жесты, вокальные жесты *А* и *О* в значении 'понимания' и 'удивления' не имеют языковых синонимов и для прояснения их значения приходится использовать тот или иной метатекст (*удивленно воскликнул, понимающе кивнул* и под.).

2.3. Вокальные жесты *А* и *О* в этих значениях сопровождаются телесными жестами «высокого передела», т. е. высокой степени условности (ср. выше, разделы 2.1.2, 2.1.5, о доминантных жестах, сопровождающих вокальный жест *А* в значении 'понимание').

3. Вокальные жесты *А* и *О* в значениях, соответственно, 'вопроса'/'пренебрежения' и 'указания', также имеют свои характеристики.

3.1. Фонематическая природа их четко определена, они не могут быть спутаны с другими вокальными жестами.

3.2. Эти вокальные жесты являются полноценными языковыми знаками, соответственно, имеют план выражения и план содержания и, как следствие, могут быть заменены синонимами, так что при их толковании не обязательно обращаться к метатексту. Так, например, вокальный жест *О* в значении 'указания' имеет близкие синонимы *во* и *вот*, вокальный жест *А* в значении 'вопроса' — синоним *Что?* (см. выше, раздел 2.1.1, п. 3), а этот же вокальный жест в значении 'пренебрежения' — синонимы (*А*) *ну его! Плевать! К черту! Что поделаешь! Ерунда!* и некоторые другие.

3.3. Как и в предыдущем случае, вокальные жесты *О* и *А* в перечисленных выше значениях сопровождаются телесными жестами высокой степени условности (о жесте *показать пальцем* как доминантном жесте для *О* указания см. [Гришина 2009], о доминантных жестах для *А* вопроса и пренебрежения см. выше, раздел 2.1.5).

Условимся называть в дальнейшем вокальные жесты первой группы ***возгласами***, второй — ***междометиями***, третьей ***частицами***. Подытожить их соотношение можно в следующей таблице:

Таблица 3

Параметры	Возгласы	Междометия	Частицы
Фонематическая природа четко определена	–	+	+
Можно заменить синонимами	–	–	+
Сопровождают телесные жесты высокой степени условности	–	+	+

Судя по значениям выбранных параметров, возгласы, строго говоря, не являются единицами языка и, соответственно, представляют собой, скорее, явления, подлежащие изучению физиологами, психологами и, в самом крайнем случае, психолингвистами. Междометия являются своего рода окультуренными, цивилизационно освоенными возгласами и в этом качестве представляют собой, безусловно, часть соответствующего языка и, соответственно, объект лингвистического анализа. Что касается частиц, то они являются полноценными языковыми единицами.

2.3.3. Имея в виду вышесказанное, мы можем обратиться к данным, содержащимся в Таблице 2.

Общее замечание: если значение данного параметра для данного класса явлений существенно отличается (причем как в большую, так и в меньшую сторону) от среднего значения данного параметра для всех рассмотренных явлений, значит, данный параметр существен для описания данного явления.

1. Группа возгласов. Для вокальных жестов *А* и *О* в этой группе характерно нисходящее движение тона (34 % против 17 % в среднем). Это объясняется тем, что реакция на физиологическое воздействие, вызвавшее эти возгласы, следует непосредственно сразу же за самим физиологическим воздействием, следовательно, вся сила звука в реакции приходится на первые миллисекунды звучания, а дальше на протяжении возгласа происходит спад звучания, который естественно сопровождается падением тона.

Отдельно следует обратить внимание на довольновысокийпроцентровногодвижениятонаувозгласов (10 %). Это объясняется тем, что для возгласа *А* в значении ‘ужас’ характерен резкий подъем тона до значений, существенно более высоких, чем средний тон данного говорящего, и удержание голоса в этом состоянии в течение относительно длительного времени.

Еще одно замечание связано с тем, что и возглас *А*, и возглас *О* могут выражать болевые ощущения говорящего. Разница между двумя этими возгласами в данном случае состоит в том, что возглас *О* имеет среди фонетических признаков огубленность, отсутствующую у возгласа *А*, соответственно, произнесение возгласа *О* требует от говорящего больших усилий, чем произнесение возгласа *А*. Как представляется, именно это является причиной того, что возглас *О* сопровождает болевые ощущения, если говорящий пытается сдерживать себя, т. е. прилагает усилия, чтобы не кричать, а возглас *А* — это выражение не сдерживаемой боли.

Заметим напоследок, что для произнесения возгласов характерен форсаж звука (по другой терминологии, — хриплый голос), т. е. включение в произнесение соответствующего гласного различного рода шумов, хрипов, щелчков и под., которые часто прерывают звучание гласного, так что иногда на фонограмме в соответствующем месте гласный как таковой не отражается, а наблюдаются только неупорядоченные шумы.

2. Группа междометий (‘удивление’ и ‘понимание’). Для междометий *А* и *О* характерно волновое движение тона, именно эта характеристика четко отличает их от частиц, ни для одной из которых волновой тон не характерен[9].

В качестве версии, объясняющей этот факт, можно предположить иконический характер данного тонового рисунка. Как известно, в русском языке стандартный вопрос связан с повышением тона (и именно этот факт определяет повышение тона на вокальном жесте *А* в значении ‘вопрос’, которое преобладает над всеми остальными способами интонировать эту частицу, что мы и наблюдаем в Таблице 2 (88 % против 22 % в среднем)). И значение ‘удивление’ (*О*), и значение понимания (*А*) включают в себя вопрос в качестве необходимой начальной составляющей (что-то типа *Что случилось?* в случае междометия *О* и *Как это понимать?* в случае междометия *А*).

При этом *О* удивления так и остается на стадии вопроса, не получая на него ответа (и, как следствие, тон 1) либо просто падает до начальной точки (восходяще-нисходящий тон, 41 % против 23 % в среднем), 2) либо фиксируется на достигнутой высоте и некоторое время длится (ровное движение тона, 16 % против 7 % в среднем)).

Что касается *А* понимания, то в этом случае ответ на заданный вопрос получен, соответственно, происходит падение тона ниже начального уровня (восходяще-нисходящий тон с падением, 45 % против 28 % в среднем), что подчеркивает нисходящую, утвердительную интонацию понимания.

3. Группа частиц. Частицы, будучи чисто языковыми единицами, вряд ли обладают иконичностью в движении тона. Скорее здесь следует ожидать некоторых факторов, связанных со структурой соответствующей языковой зоны.

О том, что *А* в значении ‘вопрос’ имеет общеязыковую вопросительную интонацию, мы уже упомянули выше. Таким образом, восходящее движение тона в данном случае предопределено значением вопросительной частицы *А*. Волновое движение тона закреплено за междометием *А*. Тем самым, в лингвистической зоне вокального жеста *А* для *А* в значении ‘пренебрежение’ из частотных движений тона остается только нисходящее, что мы и находим в таблице (55 % против 17 % в среднем).

Как видим, в зоне вокального жеста *А* нашли свои тоновые ячейки три имеющиеся языковые единицы (междометие и две частицы). В зоне вокального жеста *О* ситуация проще — там на данный момент зафиксированы две языковые единицы — междометие и указательная частица. Поскольку волновое движение тона закреплено за междометием, то на долю указательной частицы остается «неволновое» движение; «неволновых» же движений, собственно, два — нисходящее (38 % против 17 % в среднем) и восходящее (36 % против 22 % в среднем). Заметим, что эта дихотомия (волновое — неволновое движение тона) подчеркивается в случае вокального жеста *О* еще и дополнительным фонетическим различием, твердым приступом в случае указательной частицы (см. [Гришина 2009:85–87]).

---

[9] Можно ли из этого сделать вывод, что для всех первообразных вокальных междометий характерен именно волновой тоновый рисунок, и именно на основе этого признака можно отличать междометия от фонетически сходных лингвистических единиц, говорить рано за недостатком материала (остальные первообразные вокальные жесты, *У, И, Э*, гораздо менее частотны, чем *А* и *О*), однако как версию это рассматривать можно.

Напоследок заметим, что если с атрибуцией частицы *О* (указательная частица) и частицы *А* в вопросительном значении (вопросительная частица) проблем не возникает, то атрибуция частицы *А* в значении ‘пренебрежения’ неочевидна. С нашей точки зрения, эту частицу разумно считать отрицательной.

3. Заключение

В качестве заключения выскажем некоторые соображения о происхождении частиц *А* в значении ‘вопрос’ и ‘пренебрежение’ (происхождение междометий *А* и *О*, с нашей точки зрения, не подлежит лингвистическому анализу, а свои соображения о возможном происхождении указательной частицы *о* мы высказывали в работе [Гришина 2008:86–87]).

Мы уже писали выше о широко распространенном свертывании конструкций в русской устной речи. В частности, именно свертыванием вопроса до вопросительной интонации объясняется, с нашей точки зрения, вопросительная интонация у частицы *А* в значении ‘вопрос’. Вернемся еще раз к примерам (свернутый вопрос дан нижним индексом):

(1а) [Сан Саныч] *Степа!*
[Степан] *А* $_{\text{что тебе нужно}}$*?*
(2а) [Эпштейн] *Ножницы*.
[Куркова] *А* $_{\text{что ты сказал}}$*?*
(3а) [Верочка] [пугается] *А!*
[Лисичкин] *А* $_{\text{что случилось}}$*?*

Как видим, вопросы *А что тебе нужно?*, *А что ты сказал?*, *А что случилось?* являются вполне законченными высказываниями на естественном языке, что дает нам возможность предположить, что эти вопросы могут быть свернуты до начального *А*. Таким образом, в примерах 1б–3б

(1б) [Сан Саныч] *Степа!*
[Степан] *Что?* (=*Что* $_{\text{тебе нужно}}$*?*)
(2б) [Эпштейн] *Ножницы*.
[Куркова] *Что?* (=*Что* $_{\text{ты сказал}}$*?*)
(3б) [Верочка] [пугается] *А!*
[Лисичкин] *Что?* (=*Что* $_{\text{случилось}}$*?*)

мы имеем дело со свертыванием вопросов до начального *что*, а в примерах 1а–3а — со свертыванием тех же вопросов, имеющих в полном варианте начальную частицу *а*, до этого начального *а*. Заметим, что частица *ась?*, которая по смыслу полностью идентична вопросительной частице *а?* и отличается от нее лишь стилистически, в словаре М. Фасмера также возводится к частице *а* (см. [Фасмер, 1]).

Что касается отрицательной частицы *А* в значении ‘пренебрежения’, то она, с нашей точки зрения, является результатом свертывания до начального элемента конструкций типа *А ну его* (*к черту*)*!*

Эти предположения, естественно, так и останутся только предположениями, как и всякие версии о генезисе единиц, свойственных исключительно или в основном устной речи. Однако, как представляется, в современной устной речи можно найти явления, в чем-то сходные с описываемыми: сначала свертывание высказывания до начальной единицы (часто — достаточно случайной и необязательной в исходном высказывании), а затем переосмысление этой начальной единицы уже как результата свертывания.

При анализе употребления слова *да* в устной речи нами был обнаружен следующий контекст:

(21) [Костя] *Зачеркните шесть номеров. И считайте / что деньги у вас в кармане*.
[Таня] *А какие номера зачеркивать?*
[Костя] ***Да** / любые!*

Спортлото-82, Л. Гайдай, 1982

Да в последней реплике имеет все фонетические признаки *да* ‘согласия’ — паузу перед последующим текстом, ударность. При этом, однако, это *да* сопровождается нехарактерным жестом для *да* ‘согласия’ — резким движением подбородком вверх (не говоря уже о том, что трудно понять, почему в этом контексте используется любая форма согласия). Без насилия над материалом истолковать это *да* можно лишь в случае, если мы признаем, что оно является результатом свертывания до начального элемента фраз типа *Да какая разница!, Да все равно!*[10]. Косвенным подтверждением этого может послужить тот факт, что при переводе примера (21) на украинский язык, в котором, как известно, нет омонимии начальной частицы, равной по значению русскому безударному *да* (= укр. *та*), и *да* ‘согласия’ (= укр. *так*), последняя реплика в примере (21) должна быть переведена не **Так / будь-які!*, а *Та / будь-які!*

Аналогичный пример, но уже для слова *ну*:

(22) [Дятлов] *Но на работе я / ни за что!*
[Смирнов] *Угу…*
[Дятлов] *Хоть сама Софии Лорен вместо Клюевой*.
[Смирнов] *Прав! Прав-прав-прав! Скажи / а если б не на работе встретился с Клюевой?*
[Дятлов] ***Ну!** Тогда другое дело!*

Самая обаятельная и привлекательная,
Г. Бежанов, 1985

[10] Для этих фраз в качестве сопровождающих телесных жестов как раз и характерно резкое движение подбородком вперед, которое в повседневной жестикуляции замещает жест махнуть рукой в значении отбрасывания, т. е. тот же жест, который является доминантным для *А* в значении ‘пренебрежения’.

Здесь *ну* в последней реплике имеет все признаки того *ну*, которое обычно является самостоятельным высказыванием и передает общую идею побуждения, — оно является ударным, на нем происходит повышение интонации. Однако весь предшествующий контекст не поддерживает здесь значение побуждения, кроме того, это *ну* сопровождается неожиданными для *ну* ‘побуждения’ жестами — подъемом бровей, пожатием плеч, т. е. жестами дистанцирования. Единственным объяснением употребления *ну* в этом диалоге является свертывание до начальной словоформы начинающихся с безударной частицы *ну* фраз типа *Ну о чем тут говорить!, Ну что тут скажешь!, Ну что тут поделаешь!* и подобных, для которых как раз и характерны жесты дистанцирования.

Как видим, такое свертывание приводит к вынужденной ударности оставшихся словоформ и к их омонимии с полноударными *ну* и *да,* т. е. к потере связи с исходными безударными начальными частицами *ну* и *да*[11] и к их последующему переосмыслению.

Представляется, что эти два примера хорошо иллюстрируют возможность свертывания названных выше конструкций до начальной безударной частицы *а*, которая в результате получает ударение и, будучи переосмыслена, занимает собственное место в устной речи.

[11] Тот факт, что данные употребления *да* и *ну* являются результатом свертывания, подчеркивается возможностью появления в конце их произнесения гортанной смычки (*ну*ˀ, *да*ˀ), являющейся иконическим отображением резкого окончания фразы, которая должна была бы продолжаться дальше. Для стандартных *да* и *ну*, не являющихся результатом свертывания, также характерно появление в финальной части произнесения гортанной смычки, но в этом случае она имеет общее для этих единиц, а также для неформального отрицания *неа* значение вызова, фамильярности, фривольности, расслабленности (см. об этом [Кривнова 2007]).

Литература

1. *[Гольдин 1998] — В. Е. Гольдин.* Заметки о частице «вот». // Лики языка, М., 1998, с. 40–47
2. *[Гришина 2007] — Е. А. Гришина.* О маркерах разговорной речи (предварительное исследование подкорпуса кино в Национальном корпусе русского языка) // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог 2007» (Бекасово, 30 мая — 3 июня 2007 г.), с. 147–156 — Елена Гришина 1/26/09 2:28 PM (http://www.dialog-21.ru/dialog2007/materials/html/22.htm)
3. *[Гришина 2008] — Е. А. Гришина.* Частица *вот*: варианты, используемые в непринужденной речи // Инструментарий русистики: корпусные подходы (Slavica Helsingiensia 34). Хельсинки, 2008, с. 63–91 (http://docs.google.com/View?id=df52fjjj_9fntn6kxq)
4. *[Гришина 2009] — Е. А. Гришина.* К вопросу о соотношении слова и жеста (вокальный жест О в устной речи) // Компьютерная лингвистика и интеллектуальные технологии (по материалам ежегодной Международной конференции «Диалог 2009»). Вып. 8 (15) М., 2009, с. 80–90 (http://www.dialog-21.ru/dialog2009/materials/html/14.htm)
5. *[Кривнова 2007] — О. Ф. Кривнова.* Явление ларингализации в русской речи // Русский язык: исторические судьбы и современность. III Международный конгресс исследователей русского языка. М., МГУ, 2007, с. 348 (http://www.philol.msu.ru/~rlc2007/abstracts/?sectionid=12)
6. *[Ожегов, Шведова 1999] — С. И. Ожегов, Н. Ю. Шведова.* Толковый словарь русского языка. М., 1999
7. *[Ушаков 1935]* — Толковый словарь русского языка. П/р Д. Н. Ушакова. М., 1935
8. *[Шаронов 2006] — И. А. Шаронов.* О новом подходе к классификации эмоциональных междометий // Компьютерная лингвистика и интеллектуальные технологии. Труды Международной конференции «Диалог’2006» (Бекасово, 31 мая — 4 июня 2006 г.) http://www.dialog-21.ru/dialog2006/materials/html/Sharonov.htm
9. *Grishina 2007 — E. Grishina.* Text Navigators in Spoken Russian // Proceedings of the workshop “Representation of Semantic Structure of Spoken Speech” (CAEPIA’2007, Spain, 2007, 12–16.11.07, Salamanca), Salamanca, 2007, p. 39–50 (http://docs.google.com/View?id=df52fjjj_11fmxsdzdh)

Создание и лингвистическая разметка звуковой словарно-грамматической базы данных по ительменскому языку

Sound lexico-grammatical data base on itelmen language: creation and linguistic annotation

Долозова О. Н. (dolozova@gmail.com)

Санкт-Петербургский государственный университет

В докладе представлено описание звуковой базы данных по ительменскому языку, созданной на основе архивных аудиозаписей и словарных материалов. База данных включает в себя элементы лингвистической разметки, позволяющей осуществлять поиск лексем и словоформ по алфавиту, частеречной принадлежности, некоторым грамматическим и фонетическим признакам.

1. Задачи создания звуковой словарно-грамматической базы данных по ительменскому языку

Актуальность и значимость создания **электронных лингвистических ресурсов** по языкам, находящимся под угрозой исчезновения, неоднократно подчеркивалась во многих работах, посвященных этой проблематике[1]. Создание такого рода ресурсов позволяет говорить об эффективном достижении, как минимум, двух целей:

- Аккумуляция и систематизация материалов, существующих в разрозненном виде, что затрудняет доступ к ним для проведения исследований. Преобразование в электронный формат данных, существующих в рукописном виде, оцифровка аналоговых звуковых записей позволяет также сохранить эти материалы, не допустив их окончательной утраты.
- Создание дополнительной мотивации к изучению языка, благодаря представлению данных в современном компьютерном формате, обеспечивающем гибкость, доступность и привлекательность ресурса, а также за счет лингвистической классификации языковых данных по нескольким параметрам.

Наконец, поскольку речь идет о языке исчезающем[2], который не воспроизводится носителями, эти материалы представляют особую ценность. Будучи структурированы и организованы в систему, они дают возможность воссоздать утраченную на настоящий момент **звуковую материю** языка и проследить за произошедшими изменениями (звуковые записи демонстрируют аутентичное звучание речи на ительменском языке, которое современные единичные носители языка уже не способны воспроизвести). Несмотря на проблемное качество записи, эти материалы способны дать представление о звуковой форме языка и могут послужить своего рода «образцом для подражания» тем, кто будет изучать этот язык в дальнейшем.

Разработка описываемой базы данных по ительменскому языку велась в Институте филологических исследований СПбГУ в рамках проекта — Разработка национального фонда звучащей речи «Голоса народов России» при поддержке РГНФ, проект № 07-04-12163в. В ходе создания базы данных решались такие задачи как:

- перенос языковых данных на современные электронные носители, — систематизация, редактирование и структурирование языковых данных;

[1] См., например, LULCL 2005, Proceedings of the Lesser Used Languages and Computer Linguistics Conference, Bolzano, 27–28 October 2005, Ed. Izabella Ties — 336 p.

[2] По данным переписи 2002 г., ительменским языком владело 375 человек; впрочем, все они знали его значительно хуже, чем русский. По другим данным, уже в 1989 году языком владело менее 100 человек. (http://lingsib.iea.ras.ru/ru/languages/itelmen.shtml)

- разработка формата представления и способов подачи языкового материала;
- конвертация в формат базы данных имеющейся структурированной информации;
- разработка удобного и интуитивно понятного интерфейса, позволяющего осуществлять быстрый поиск необходимой информации по базе данных.

Материалом для описываемой звуковой словарно-грамматической базы данных послужили:

1) ительменские архивные полевые тетради[3], представляющие собой записи словарной программы;
2) соответствующие тетрадям звуковые записи, представленные на аналоговых носителях, которые в ходе выполнения проекта были оцифрованы и внесены в электронный каталог.

Языковые материалы этих двух типов были систематизированы, звуковые записи словаря соотнесены с расшифровками и транскрипциями. В качестве **программной среды** для реализации был выбран формат базы данных **Microsoft Access**, позволяющий реализовать поиск по различным параметрам в структуре реляционного типа. Этот формат также является достаточно гибким с точки зрения возможностей редактирования и пополнения базы новыми языковыми данными.

Говоря о выборе формата представления данных, следует заметить, что в мировой практике документирования и обработки языкового материала, сопровождаемого аудио или видеозаписью, сложились определенные стандарты. Имеется специализированный инструментарий, широко используемый и хорошо зарекомендовавший себя, в частности, такие программы как Toolbox[4] и ELAN[5]. Впрочем, как отмечается в статье J. Good[6], универсальных рекомендаций о том, какое программное обеспечение следует использовать при работе с тем или иным типом языковых данных, не существует. В нашем случае использование этого инструментария не позволило бы эффективно решить весь спектр поставленных задач, или создавало бы дополнительные сложности для потенциальных пользователей. Причины, по которым был выбран универсальный формат баз данных Microsoft Access, а не специализированные лингвистически ориентированные инструменты Toolbox и ELAN, перечислены ниже.

- В создаваемой базе данных исходной единицей хранения информации являются лексемы и соотносимые с ними производные словоформы, которые характеризуются по заданному набору признаков. Такой инструментарий как Toolbox и ELAN, хотя и позволяет создавать, в том числе, словарные материалы, однако в первую очередь ориентирован на обработку текстов как линейно разворачивающейся во времени звучащей цепи (ELAN) и последующий анализ составляющих элементов (ELAN, Toolbox). В нашем случае сегментация словоформ не осуществляется, все классификации оперируют словоформами как целостными единицами.
- Набор признаков, релевантных для осуществления поиска в создаваемой базе данных, и способ их организации предполагает особый тип наглядного представления данных, а также возможность соотнесения единицы определенного типа с несколькими единицами другого типа (например, наличие нескольких вариантов произнесения той или иной словоформы). Таким образом, более востребованным оказывается иерархический способ организации данных, а не последовательно линейный.
- Система поисковых запросов организована с учетом двух типов адресата: лингвисты-исследователи (на них ориентирована более специальная лингвистическая разметка) и представители языкового сообщества, те, кто с помощью базы данных сможет изучать язык (для них созданы интуитивно понятные поисковые запросы, содержащие общеизвестную лингвистическую терминологию — например, существительное, словоформа).
- Более широкое распространение формата Microsoft Access в нашей стране, более простой интерфейс, не требующий длительного специального изучения, как в случае с программами ELAN и Toolbox, которые имеют достаточно сложную структуру и ориентированы на более подготовленных пользователей компьютера.

2. Структура создаваемой базы данных

Итак, как уже было отмечено ранее, ключевыми **единицами хранения информации** в нашей базе данных являются **словоформы** и **лексемы.**

Лексемы репрезентированы исходными словарными формами лексических единиц[7]. В базе дан-

---

[3] Записи осуществлялись в ходе лингвистической полевой экспедиции в сентябре-октябре 1984 г. в поселке Ковран. Исследователи — д. филол. н., проф. А. П. Володин, д. филол. н. А. С. Асиновский

[4] http://www.sil.org/computing/toolbox/

[5] http://www.lat-mpi.eu/tools/elan/

[6] Jeff Good, Data and language documentation. To appear in Peter Austin and Julia Sallabank (eds.), Handbook of Endangered Languages. Cambridge: Cambridge University Press. http://www.acsu.buffalo.edu/~jcgood/publications.html

[7] В некоторых случаях для глагольных форм (особенно тех, которым в переводе соответствует целое словосочетание) в качестве исходной выступает не форма инфинитива, а личная форма (как правило, 1-е или 3-е лицо настоящего времени)

ных они соотнесены со всеми встретившимися в материалах соответствующими производными формами. Таким образом, в качестве единицы, по которой осуществляется поиск, может выступать: — *лексема*: в этом случае в ответ на поисковый запрос мы получаем список всех соотнесенных с ней производных словоформ, имеющихся в БД; — *словоформа*: в ответ на поисковый запрос мы получаем подробную информацию об интересующей нас словоформе и ссылку на соответствующую ей лексему.

Специфика экспедиционной программы, в соответствии с которой осуществлялась запись, заключалась в том, что конкретные грамматические формы фиксировались лишь выборочно. При подготовке базы данных в ряде случаев были добавлены исходные формы, которых не было в расшифровках и звуковых материалах. Эти формы не сопровождаются звучанием и транскрипцией, но являются полноценной единицей хранения информации, обеспечивающей более удобный систематизированный поиск.

Каждая **словоформа** соотнесена с **переводным эквивалентом** на русском языке и охарактеризована по нескольким **грамматическим и фонетическим признакам**.

В базе данных реализовано **5 основных типов поисковых запросов,** которые в обобщенном виде релевантны для любого языка и могут быть конкретизированы в зависимости от того, какие **грамматические и фонетические особенности** представляют интерес. Поисковые запросы в базе данных основаны на предварительно осуществленной разметке, представленной в электронных таблицах, на основе которых была произведена конвертация в формат базы данных.

2.1. Поиск всех лексем (словоформ)

Создан поисковый запрос, по которому может быть получен упорядоченный по алфавиту список всех лексем или всех словоформ, имеющихся в базе данных. При этом для каждой лексемы предусмотрены:

- возможность просмотра всех соответствующих ей словоформ, сопровождаемых переводом на русский язык и образцами звучания (рис. 1);
- возможность перехода к подробному описанию каждой словоформы, включающему в себя перевод, транскрипцию, образцы звучания, грамматические и фонетические признаки, лексический комментарий, возможные варианты орфографической записи, метаданные (номер записи, шифр звукового файла, код информанта) (рис. 2).

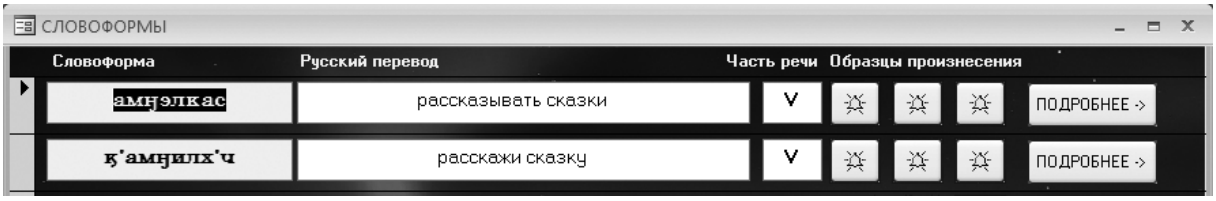


Рис. 1. Просмотр всех словоформ для заданной лексемы

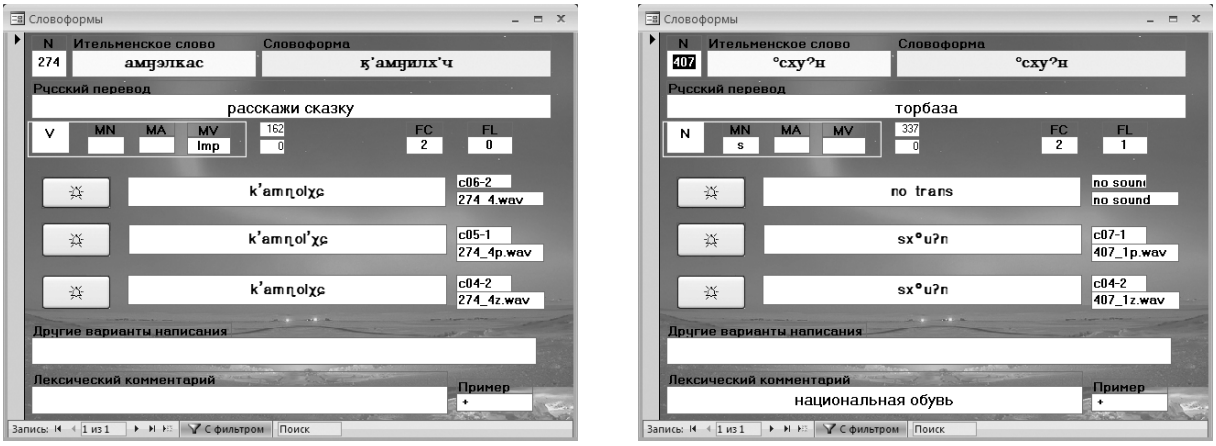


Рис. 2. Примеры описания словоформ по заданным в базе данных параметрам

2.2. Поиск всех лексем (словоформ), начинающихся на определенную букву ительменского алфавита

Результат поиска — список лексем (словоформ) на заданную букву, имеющихся в базе данных. Поиск реализован благодаря представлению списков лексем и словоформ в алфавитном порядке. В базе данных представлена не вся алфавитная последовательность, а только те символы, которым соответствует информация (рис. 3).

2.3. Поиск словоформ по признаку принадлежности к определенной части речи

Поисковый запрос в базе данных построен таким образом, что в виде отдельных списков могут быть получены все словоформы существительных, прилагательных, глаголов, наречий, местоимений.

2.4. Поиск определенных грамматических форм (типов словоформ)

Осуществляется за счет указания дополнительных грамматических признаков для различных частей речи. В базе данных могут быть получены списки: существительных единственного числа, существительных множественного числа, существительных в форме диминутива, существительных в формах косвенных падежей, существительных собирательных, глагольных форм императива, форм суперлатива для прилагательных (рис. 4).

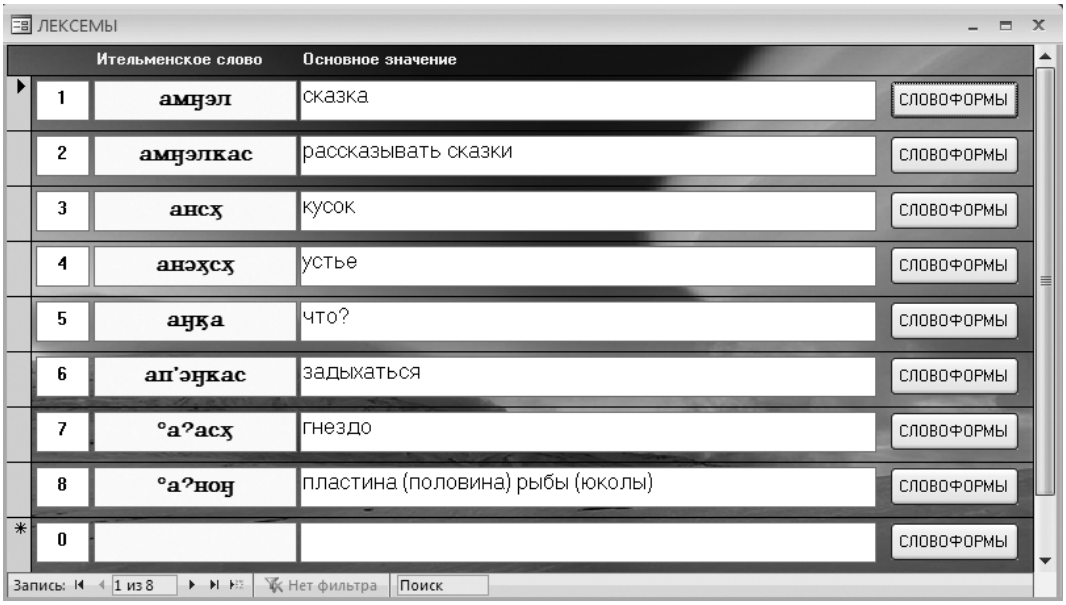


Рис. 3. Просмотр всех лексем на заданную букву алфавита

Рис. 4. Окно поиска по частям речи и типам словоформ

2.5. Поиск по фонетическим характеристикам

В базе данных предусмотрена возможность поиска по двум видам фонетических характеристик: наличие/отсутствие лабиализации («огубленности»), а также количество согласных в консонантных сочетаниях. Таким образом, может быть получен список словоформ, характеризующихся наличием гармонии лабиализации (рис. 5).

В виде отдельного списка, упорядоченного по алфавиту, могут быть представлены словоформы, содержащие консонантные сочетания, включающие от 3 до 6 согласных (рис. 6).

Звуковая составляющая базы данных организована следующим образом: каждой словоформе соответствует от одного до трех звуковых файлов, репрезентирующих произнесение различных информантов — носителей ительменского языка.

Всего в БД представлены записи от 3 информантов — все они — жители одного населенного пункта, примерно одного возраста, среди них — 2 женщины и 1 мужчина, поэтому представленная вариативность носит, скорее, идиолектный характер. В исходных записях словоформа в исполнении одного диктора произносилась 2–3 раза — в базе данных этот принцип сохранен, поскольку некоторые повторы звучат более четко.

Каждому варианту звучания соответствует *реальная фонетическая транскрипция*, записанная в символах IPA (международного фонетического алфавита). В базу данных были внесены варианты

Рис. 5. Фрагмент списка словоформ, характеризующихся наличием лабиализации

Рис. 6. Фрагмент списка словоформ, содержащих консонантные сочетания, состоящие из 3 согласных и более

транскрипции, зафиксированные исследователями в ходе экспедиционной работы, с некоторыми уточнениями и унификацией символов.

Представленные транскрипции могут стать предметом отдельного более детального исследования, в результате которого они могут подвергнуться корректировке. К сожалению, возможности корректировки ограничены низким качеством записи. Для того чтобы внесение уточнений в транскрипцию стало возможным, необходимо осуществить дополнительную реставрацию записи.

Орфографическая запись, представленная в таблице, максимально приближена к современной орфографической норме ительменского языка[8]. Впрочем, в ряде случаев представлены также варианты орфографической фиксации, которые могут более точно отражать те или иные произносительные особенности. Эти варианты внесены в специально созданное в БД поле «Другие варианты написания», предназначенное для внесения комментариев. В поле «Лексический комментарий» представлены некоторые пояснения к значению, комментируются специфические реалии.

Интерфейс базы данных интуитивно понятен и реализован на русском языке. Разработана удобная система непротиворечивой **аннотации** для осуществления поиска в базе данных. В большинстве своем названия полей интуитивно понятны любому пользователю системы, часть из них приводится на русском языке: *ительменское слово, основное значение, словоформа, русский перевод, часть речи, образцы произнесения, другие варианты написания*, *лексический комментарий*. Для некоторых полей, содержащих более специальную грамматическую и фонетическую информацию, использовались условные обозначения в символах латиницы.

В создаваемой базе данных исследователям открыт доступ к редактированию, добавлению и удалению языковых данных, а также изменению форм и типов поисковых запросов. Общая схема аннотации и структура базы данных может стать основой для организации материала по другим языкам, а типы созданных поисковых запросов могут быть конкретизированы в соответствии со спецификой описываемого языка.

---

[8] Эта норма отражена, в частности, в издании Володин А. П., Халоймова К. Н. Словарь ительменско-русский и русско-ительменский: Пособие для уч-ся нач. шк. — Л., 1989. — 255 с.

Итеративное применение алгоритмов снятия частеречной омонимии в русском тексте[1]

Iterative application of part-of-speech disambiguation algorithms for russian text

Епифанов М. Е. (xeme@rambler.ru),
Антонова А. Ю. (a-antonova@list.ru),
Баталина А. М. (batalina_anna@rambler.ru),
Кобзарева Т. Ю. (stamstam@mtu-net.ru),
Лахути Д. Г. (delir1@yandex.ru)

Российский государственный гуманитарный университет, Москва

В статье дается краткое обоснование итерационной версии блока алгоритмов анализа частеречной омонимии в русском предложении и описывается соответствующее расширение функциональности инструментальной среды для экспериментов с алгоритмами синтаксического анализа (ЭСЛА).

Введение

Описываемая работа представляет собой один из этапов исследования и реализации возможности автоматизации поверхностно-синтаксического анализа русского текста средствами, использующими только грамматические (морфосинтаксические) характеристики входящих в него слов. Постановка задачи и предварительные результаты этого исследования описывались в материалах конференции ДИАЛОГ ([1]–[3]) и в указанных там источниках.

Разрабатываемая система анализа состоит из семи основных блоков:

- постморфологии (постморфологического анализа), в котором исправляются некоторые ошибки морфологического анализа, обрабатываются несловарные собственные имена и числовые (количественные) группы;
- снятия частеречной омонимии;
- предсегментации — построения некоторых связей: сложного сказуемого, определительных атрибутивных отношений именных групп, предложных групп и т. д., задающих границы проективных фрагментов, минимально необходимых для сегментации;
- сегментации (сегментационного анализа): построение сегментов — простых главных и придаточных предложений, деепричастных и других обособляемых оборотов;
- внутрисегментного анализа — поиск связей слов внутри построенных сегментов;
- межсегментного анализа — построение связей сегментов;
- построения субграфа кореференции.

В своей работе мы исходили из того, что «сколь бы глубоко и тщательно ни разрабатывался синтаксический анализатор системы МП, в нем неизбежно остаются белые пятна и лакуны. Многие лингвистические факты, в том числе критически существенные для работы анализатора, никогда не попадали в поле зрения ученых просто потому, что они не имели возможности в массовом порядке оперировать материалом неверного или неожиданного синтаксического разбора предложений. Именно такой материал в изобилии поставляет развитая система автоматической обработки текстов, а работа с этим материалом позволяет выявить лакуны научного описания и устранить их» [4]. Это значит, что внесение изменений в правила (алгоритмы) работы системы синтаксического анализа должно быть регулярным режимом ее работы. Для того, чтобы это не приводило каждый раз к перепрограммированию всей системы или значительных ее частей, была разработана специальная программная среда — экспе-

[1] Работа выполнена при частичной поддержке гранта РФФИ № 09-06-00275-а.

риментальная система работы с лингвистическими алгоритмами (ЭСЛА) [5].

Предметом описываемой работы является блок снятия частеречной омонимии[2]. Мы хотим подчеркнуть, что речь идет не о конкретном, тем более не о *реализованном* блоке (системе алгоритмов) снятия частеречной омонимии (создание которого является одной из *целей* проводимой нами работы), а об *инструменте*, предназначенном для использования при выборе оптимального варианта такого блока (см. [2]).

Блок снятия частеречной омонимии и его повторное применение в процессе анализа

На предшествующих этапах исследования было выделено всего 63 разных типа частеречной омонимии [6]; из них было выделено 15 наиболее часто встречающихся, для которых в экспериментальном режиме были запрограммированы соответствующие модули снятия (распознания) омонимии.

В актуальной версии модули снятия омонимии разных типов работают последовательно в порядке, задаваемом экспериментатором (т. е. сначала предложение обрабатывается модулем снятия омонимии одного типа, затем модулем снятия омонимии другого типа и т. д.). В ходе предшествующих исследований был определен «условно оптимальный» порядок применения 15 модулей снятия омонимии: сложные предлоги, сложные наречия, словосочетания с частицами *не* и *ни* (типа «ни у кого», «ни для чего»), фразеологизмы, приравниваемые к одному слову), вводные слова и конструкции[3], краткое прилагательное — существительное, деепричастие — существительное, существительное — глагол в спрягаемой форме, существительное — прилагательное, предлог — существительное, краткое прилагательное — существительное — списочный глагол, «его — ее — их», существительное в творит. падеже — наречие, морфологически уникальные, но часто встречающиеся омонимы (типа «том», «отчего», «три» и т. п.).

Действие данного блока основано на анализе достаточно близкого синтаксического контекста, причем до собственно синтаксического анализа. При этом слова, сами являющиеся частеречными омонимами, не могут использоваться при анализе, так что при наличии таких слов в контексте обрабатываемого омонима некоторые диагностические конфигурации, используемые в алгоритмах, не будут работать. Поэтому при наличии в предложении более чем одного омонима при однократном последовательном применении указанных модулей некоторые омонимы могут остаться нераспознанными, но будут распознаны при следующем применении (следующих применениях) модулей снятия омонимии.

Рассмотрим примеры, для успешного анализа которых необходимо повторное применение модулей снятия омонимии.

(20) Омонимия «краткое прилагательное (Abr) — существительное (N)»:

*Она не помнит ни **добра**, ни **зла**.*

Омонимы данного типа распознаются справа налево. При первом проходе снимается омонимия ***добра***=N; омонимия ***зла*** остается нераспознанной. При втором проходе определяется ***зла***=N.

(21) Омонимия «(местоименное) прилагательное (A) — (местоименное) существительное (N)», «(местоименное) прилагательное (A) — (местоименное) существительное (N) — существительное (N)» («уникальная»), «краткое прилагательное (Abr) — наречие (D) — списочный предикатив (Vcп)»:

***Все это** говорит о **том**, что емкость рынка **достаточно** велика*.

В соответствующих алгоритмах первый из этих типов омонимии распознается справа налево, второй и третий — слева направо. При первом проходе распознаются все омонимы, кроме ***это***, который распознается при втором проходе.

Говоря в общем, встречаются ситуации, когда в предложении содержится несколько омонимов, причем некоторые их пары зависимы в смысле снятия омонимии: чтобы разрешить омонимию типа **A** для слова **a** требуется предварительно снять омонимию типа **B** для слова **b**. Однако в рамках однократного применения блока снятия омонимии слово **b** анализируется после слова **a**. Если удастся разрешить омонимию для **b** при некотором применении блока снятия омонимии к предложению, то для **a** она снимется при следующем проходе.

Повторное применение блока снятия омонимии снижает (вычислительную) эффективность всей процедуры синтаксического анализа. Разработанная система лингвистических алгоритмов Т. Ю. Кобзаревой — открытая система. В процессе ее совершенствования предполагается ее расширение, это относится и к этапу анализа омонимии: в уже имеющиеся алгоритмы могут добавляться новые правила, что увеличит их размер; анализ пока не рассматриваемых типов омонимии потребует добавления новых алгоритмов (как указывалось выше, сейчас рассматривается 15 типов из 63-х).

---

[2] Далее под омонимией будет везде пониматься частеречная омонимия (омонимия частей речи).

[3] Перечисленные типы рассматриваются как случаи омонимии частей речи.

Но, быть может, изменение организации блока и перестройка его алгоритмов позволит избежать повторных проходов?

К сожалению, ни положительного, ни отрицательного «гарантированного» ответа на этот вопрос не существует. Даже если бы мы (путем значительных усилий) добились такого результата, исходя из некоторого состояния подсистемы алгоритмов на фиксированном множестве рассмотренных примеров с подобной зависимостью снятия омонимии, то мы все равно не смогли бы гарантировать успех однократного применения перестроенного таким образом блока на все новых и новых примерах. К тому же открытость системы алгоритмов потребует новых перестроек при ее расширениях. Интуитивно ясно, что разработать «однократно применяемый» блок (вообще говоря, расширяемый) при всех возможных случаях зависимости снятия одной омонимии от успешности разрешения другой, вряд ли удастся. Доказать же невозможность подобного блока, скорее всего, *невозможно*[4].

При исследовании этого вопроса можно ограничиться только рассмотрением возможности изменения порядка, в котором анализируются «зависимые» омонимы **a** и **b**.

В рассматриваемой подсистеме каждый тип снятия омонимии обрабатывается своим алгоритмом.

Если типы **A** и **B** омонимов **a** и **b** различны, как в примере (2) выше, то порядок их анализа обусловлен порядком применения алгоритмов снятия омонимии этих типов. В описанной выше ситуации (снятие омонимии **A** обусловлено снятием омонимии **B**, но **a** анализируется до **b**) изменение порядка применения алгоритмов при их однократном использовании в процессе анализа «спасет ситуацию» для данного предложения и ему (в этом смысле) подобных в случае, если **B** может быть разрешена независимо от **A**. Но тогда перестанут успешно анализироваться предложения, в которых, наоборот, для разрешения омонимии **B** требуется, чтобы уже была разрешена омонимия **A**. Наличие примеров предложений для обоих порядков применения алгоритмов служит (при данном ограничении) аргументом для повторного применения модулей в блоке снятия омонимии.

В основе составляющих данный блок алгоритмов лежит цикл прохода по словам-омонимам предложения в прямом (слева направо) либо обратном порядках. На каждом шаге этого цикла для обрабатываемого омонима анализируются контексты с целью выявления возможности снять омонимию. Если **A=B**, то последовательность рассмотрения слов **a** и **b** определяется направлением такого прохода, как в примере (1) — справа налево. В алгоритме правила анализа контекстов рассматриваемого слова, вообще говоря, сформулированы с учетом направления прохода. В принципе их можно изменить в соответствии с изменением на противоположное направления прохода, однако в случаях, когда они уже достаточно хорошо отлажены, или некоторые из них могут усложниться, делать это нежелательно. Кроме того, изменение направления прохода может привести к тому, что повторное применение блока теперь потребуется для других предложений, содержащих другую зависимость между омонимами типа **A**, рассмотренную в других правилах алгоритма.

Заметим, наконец, что наличие в предложении нескольких подобных зависимостей может привести к необходимости повторять применение модулей снятия омонимии несколько раз.

Итеративная версия блока снятия частеречной омонимии

Универсальным и наиболее простым подходом к реализации повторов анализа омонимии является итеративная версия блока, в которой все составляющие его алгоритмы применяются однократно в определенном порядке при каждой итерации. Схематично основной цикл блока можно описать как

```
while Not IsEqual(S_curr, S_prev)do
  begin
    S_prev:=Copy(S_curr);
    применить последовательно модули
    снятия омонимии к S_curr
  end
```

Цикл выполняется, пока применение алгоритмов «плодотворно», т. е. удается снять хотя бы одну какую-либо омонимию. Как только состояние объектной модели анализируемого предложения после очередной итерации S_{curr} совпадает с таким состоянием после предыдущей итерации S_{prev} («новых изменений нет»), цикл прекращается, переходим к следующей стадии синтаксического анализа.

В рассмотренной в предыдущем пункте ситуации, если на некотором шаге цикла будет снята омонимия **B**, то на следующем шаге разрешится и омонимия **A**.

Итеративная версия блока программно реализована, как и другие лингвистические алгоритмы, на входном языке системы ЭСЛА. На уровне ядра

[4] Приведенные здесь кратко рассуждения аналогичны рассуждениям, мотивирующим построение теории NP-полных задач (см., например, Гэри М., Джонсон Д. Вычислительные машины и труднорешаемые задачи // М.: Мир, 1982). Однако, ввиду открытости самой системы алгоритмов и сложности ограничения (фиксации) случаев зависимости снятия одной омонимии от успешности снятия другой, не удается адекватно сформулировать «практически осмысленную» задачу, для которой можно было бы доказать ее NP-полноту.

интерпретатора ЭСЛА реализованы, как системно-определенные «примитивы», функции сохранения состояния объектной модели анализируемого предложения и сравнения двух состояний объектных моделей.

По мнению авторов, применение итеративного блока снятия омонимии делает принципиально несущественным порядок анализа омонимов в предложении.

Заметим, что в приведенном здесь цикле последняя итерация всегда осуществляется «вхолостую», она служит лишь основанием для того, чтобы, убедившись в отсутствии вновь разрешенных омонимов, завершить выполнение блока. Так, в примерах (1) и (2) цикл будет прекращен лишь при проверке условия в начале четвертого шага. В случаях, когда в предложении успешно снимаются все омонимии анализируемых алгоритмами блока типов, последняя итерация выполняется быстрее предыдущих. Это связано с тем, что на последнем шаге цикла каждый модуль не находит омонимов «своего» типа и сразу прекращает работу.

Выше уже отмечалось, что повторное применение алгоритмов снятия омонимии снижает эффективность всего синтаксического анализа. Поэтому актуально выявление статистически значимых возможностей такого изменения порядка анализа омонимов, после которого для довольно часто встречающейся зависимости снятия омонимии количество итераций уменьшается.

Другим способом осуществить подобную оптимизацию является непосредственный вызов одного алгоритма из другого. Точнее, если снятие омонимии **A** обусловлено снятием омонимии **B** и в анализируемом (в алгоритме снятия **A**) контексте встретился омоним типа **B**, то к нему прямо здесь же и применяется модуль для **B**. Однако такие рекурсивные вызовы одного алгоритма из другого усложняют алгоритмы (увеличивается количество ветвлений) и тем самым затрудняют их последующую модификацию и отладку.

Расширение функциональности инструментальной среды

С целью более удобного проведения экспериментов с алгоритмами анализа частеречной омонимии и их отладки разработана форма, специализированная на совершенствование рассматриваемого здесь блока. В ней пользователь может

- выбрать подмножество модулей, которые он хочет применить в процессе анализа тестового предложения. Эти актуальные модули выбираются не только из тех, что составляют блок снятия омонимии, но и из входящих в блок постморфологии, т. к. последний применяется до анализа омонимии. Применение только части модулей, и, в частности, единственного алгоритма, может существенно упростить исследование поведения реализованных в узлах алгоритмов правил, обеспечивающих распознавание актуальных для анализа данного предложения ситуаций, в том числе возможности снятия омонимии в конкретных контекстах. Эта же функциональная возможность при отладке облегчает локализацию ошибок программной реализации алгоритмов и неточностей в них самих.
- задать порядок применения выбранных модулей, что позволяет оптимизировать структуру блока. Список выбранных пользователем модулей в заданном порядке сохраняется в виде отдельного файла.
- указать, применять ли выбранные модули итеративно или однократно.
- просмотреть результат анализа предложения непосредственно в одном из полей формы и сохранить его в виде отдельного файла.
- просмотреть *трассу прохождения* — последовательность действительно посещенных узлов применяемых алгоритмов в процессе анализа предложения непосредственно в одном из полей формы. Эта трасса также сохраняется в отдельном файле-журнале.

На основании сформированных пользователем данных (выбор модулей и порядок их применения, указание, применять ли их итеративно) генерируется так называемый *основной*[5] (головной) алгоритм на входном языке системы ЭСЛА, который и применяется к тестовому предложению. В основном алгоритме осуществляются вызовы модулей из заданной их последовательности, причем такая последовательность вычисляется либо итеративно, либо однократно. Код основного алгоритма сохраняется в отдельном файле и, при необходимости, может быть использован для пошаговой отладки в интерпретаторе ЭСЛА [3].

Все перечисленные выше *выходные* файлы помещаются в папку так называемого *проекта выполнения* (прогона, runtime project) сгенерированного основного алгоритма, созданную в процессе его применения к тестовому примеру. Туда же будут положены и другие, связанные с данным прогоном алгоритма *исходные*, *промежуточные* и выходные файлы:

- с текстом тестового примера,
- с результатом морфологического анализа предложения,
- с исходным состоянием объектной модели предложения, представленном во входном языке ЭСЛА и полученном соответствующей конверсией результата морфоанализа,

[5] Аналог основной функции или процедуры в процедурных языках программирования, например функции main() в C или C++.

- xml-файлы с данными для показа результата синтаксического анализа в одной из двух предназначенных для этого программ (разработанная в последнее время интерактивная подсистема визуализации результата синтаксического анализа обладает существенно более широкими функциональными возможностями, в ней удобно просматривать графы синтаксического разбора больших предложений, в то время как прежде использовавшийся модуль и сейчас часто применяется для «быстрой» оценки правильности результата).

Заключение

Реализация и последующее применение итеративной версии блока снятия частеречной омонимии расширило класс успешно анализируемых предложений.

В настоящее время проводится отладка блока, проводятся эксперименты с целью оптимизации его структуры. Эту работу существенно облегчает специально разработанный инструмент, позволяющий применять алгоритмы блока к тестовому предложению избирательно и в нужном порядке.

Литература

1. *Кобзарева Т. Ю., Лахути Д. Г., Ножов И. М.* Модель сегментации русского предложения. // Диалог'2001. Аксаково 2001. Т. 2, с. 185–194.
2. *Баталина А. М., Айриян Г. Ю., Епифанов М. Е., Кобзарева Т. Ю., Лахути Д. Г.* Автоматизация отладки алгоритмов поверхностно-синтаксического анализа // Труды международной конференции «Диалог 2005». М.: Наука, 2005. С. 45–50.
3. *Баталина А. М., Епифанов М. Е., Кобзарева Т. Ю., Кушнарёва Е. В., Лахути Д. Г.* Экспериментальная реализация сегментационного анализа русского предложения // Тр. Междунар. конф. «Диалог 2007». М. РГГУ, 2007. С. 29–34.
4. *Иомдин Л. Л.* Уроки русско-английского (из опыта работы системы машинного перевода). // Компьютерная лингвистика и интеллектуальные технологии: Труды Международной конференции Диалог'2002. Т. 2. С. 234–244.
5. *Баталина А. М., Епифанов М. Е., Ивличева О. О., Кобзарева Т. Ю., Лахути Д. Г.* Инструментальная среда для экспериментов с алгоритмами поверхностно-синтаксического анализа // Компьютерная лингвистика и интеллектуальные технологии: Труды Международной конференции «Диалог 2004», М.: Наука, 2004. С. 32–38.
6. *Кобзарева Т. Ю., Афанасьев Р. Н.* Универсальный модуль предсинтаксического анализа омонимии частей речи в РЯ на основе словаря диагностических ситуаций // Компьютерная лингвистика и интеллектуальные технологии. Тр. Междунар. сем. Диалог 2002. Т. 2. М., 2002. С. 258–268.

Терминологический анализ текста на основе лексико-синтаксических шаблонов

Analysis of text terminology based on lexicosyntactic patterns

Ефремова Н. Э. (nvasil@list.ru),
Большакова Е. И. (bolsh@cs.msu.ru),
Носков А. А. (alexey.noskov@gmail.com),
Антонов В. Ю. (avadim@gmail.com)

МГУ им. М. В. Ломоносова, факультет ВМиК

Характеризуются лексико-синтаксические шаблоны, специфицирующие особенности терминоупотребления в научно-технических текстах на русском языке. Описываются результаты экспериментального исследования основанных на этих шаблонах процедур автоматического выявления терминов в текстах.

1. Введение

Проблема автоматического выявления в тексте на естественном языке терминов — слов и словосочетаний, называющих понятия определенной проблемной области (ПО), изучается с точки зрения разных приложений, таких как реферирование и аннотирование тестов, извлечение знаний из текстовых источников, создание терминологических словарей, тезаурусов и онтологий ПО. Общим при решении этой проблемы является применение частичного синтаксического анализа текста и опора при распознавании терминов на лингвистические и статистические критерии [1, 2]. Статистические критерии так или иначе основаны на частоте употребления терминов и дают приемлемую точность и полноту распознавания только на корпусах текстов. Лингвистические критерии учитывают в первую очередь типичную структуру именных терминологических словосочетаний (частеречную принадлежность слов, входящих в термин, и синтаксические связи между ними), и реже — контексты употребления терминов, свойственные конкретной узкой ПО. Ряд исследований, проведенных для распознавания терминологических словосочетаний английского и французского языков (в частности, [3]) показывает, что более полное использование информации о лексических и синтаксических особенностях терминоупотреблений в обрабатываемых текстах повышает точность и полноту автоматического распознавания терминов.

Терминологический анализ текста, необходимый в таких приложениях, как литературно-научное редактирование и перевод текстов с одного языка на другой, предполагает распознавание в тексте различных терминоупотреблений, причем как можно более полное (при приемлемой точности). В настоящей работе описывается применяемый для решения этой задачи единый подход к представлению необходимой разнородной лингвистической информации в виде *лексико-синтаксических шаблонов* языка LSPL. Этот язык был предложен в работе [4], и для него были реализованы программные средства автоматического распознавания в тексте на русском языке конструкций по заданному шаблону [5].

Поскольку наиболее характерными в плане терминологии являются тексты научно-технического стиля, именно к ним в первую очередь и применялся указанный подход. В результате изучения различных терминологических словарей (более 15 000 словарных статей) и научно-технических текстов на русском языке (около 330 текстов) из области физики и информатики была проведена формализация лексико-синтаксической информации, необходимой для распознавания терминоупотреблений, и создан представительный набор LSPL-шаблонов. Кроме шаблонов, описывающих структуру терминологических словосочетаний, в набор входят шаблоны характерных конструкций определения новых, *авторских* терминов (например: *Под конвейерным режимом понимают такой вид обработки…*). Поскольку часто анализ текста проис-

ходит с привлечением терминологического словаря, зафиксированные в нем *словарные* термины также представляются в виде шаблонов. Кроме того, набор включает шаблоны *текстовых вариантов* терминов [6], учитывающих возможное варьирование в тексте одного термина (*текстовая коллекция — коллекция текстов — коллекция*) и **соединение** в тексте нескольких терминологических словосочетаний, при которых соединяемые сочетания могут разрываться (*входные и выходные данные — входные данные* и *выходные данные*).

В работе кратко рассматриваются средства языка LSPL, необходимые для формализации терминоупотреблений. Характеризуется созданный набор LSPL-шаблонов и базирующиеся на нем процедуры автоматического выделения в научно-техническом тексте различных терминоупотреблений, в том числе — разрывных. Приводятся результаты экспериментального тестирования этих процедур, и формулируется стратегия их совместного применения, позволяющая в целом улучшить показатели распознавания терминов в тексте (F-меру).

2. Формализация терминоупотреблений на основе языка LSPL

В общем случае лексико-синтаксический шаблон языка LSPL задает последовательность *элементов-слов*, из которых должна состоять описываемая языковая конструкция, и указывает условия синтаксического согласования этих элементов. Например, шаблон `A N <A=N>` описывает словосочетания из прилагательного `A` и существительного `N`, согласованных по их общим морфологическим характеристикам: *логический вывод*, *реактивной силы*. Для элементов-слов могут быть заданы не только часть речи, но и конкретизированы их отдельные морфологические характеристики (род, число, падеж и др.), а также — лексема, например, шаблон `N <базис,N=sing>` описывает все формы единственного числа слова *базис*.

Для описания конкретных словоформ и символов, встречающихся в описываемой конструкции, в шаблоне используется *элемент-строка*, например: `"базисом"`.

Язык шаблонов является достаточно мощным, в шаблоне могут задаваться такие сложные элементы, как:

- *повторения элементов*, задаваемые в фигурных скобках с указанием количества повторений;
- *опциональные элементы*, записываемые в квадратных скобках;
- *альтернативные элементы*, указываемые через символ |.

Например, шаблон `{A} N1 [N2<c=gen>] <A=N1>` определяет именную группу из нескольких прилагательных, согласованного с ними существительного и опционального существительного в родительном падеже (*немаркированный квантор общности, двойной электрический слой*). Заметим, что для обозначения в шаблоне разных элементов одной части речи (существительных `N1` и `N2`) используются числовые индексы.

Язык предоставляет возможность давать имена шаблонам и устанавливать у шаблонов параметры, что позволяет использовать для описания сложных конструкций уже определенные шаблоны (с конкретизацией при необходимости морфологических характеристик входящих в них слов-элементов). Например, задав имя `NP` для вышерассмотренного шаблона и установив его параметрами морфологические характеристики входящего в него существительного:

```
NP = {A} N1 [N2<c=gen>] <A=N1> (N1),
```

можно описать шаблон именной группы, стоящей в творительном падеже: `NP<c=ins>`.

Дополнительно, кроме условий согласования элементов, в шаблоне могут быть заданы *словарные условия*: запись `<Syn(N2,N4)>` означает, что существительные `N2` и `N4` являются синонимами, зафиксированными в соответствующем словаре.

Рассмотренные средства языка шаблонов LSPL позволили описать:

1. морфосинтаксические образцы терминологических слов и словосочетаний;
2. типичные конструкции определения авторских терминов;
3. характерные контексты введения синонимов терминов;
4. входы используемого при распознавании терминов терминологического словаря;
5. правила образования лексико-синтаксических вариантов терминов;
6. правила образования в тексте соединений нескольких терминологических словосочетаний.

Примеры шаблонов каждой группы представлены в Табл. 1; ее последний столбец содержит соответствующие примеры терминов и их употреблений.

В первую группу шаблонов включены 7 наиболее частотных морфосинтаксических образцов терминологических слов и двух- и трехсловных сочетаний. Каждый шаблон фиксирует часть речи входящих в термин слов и их морфологические характеристики.

Вторая группа шаблонов получена в ходе формализации характерных для научной прозы конструкций, используемых при введении в текст новых, *авторских* терминов, например: *Слабовзаимодействующие массивные частицы назовем **вимпами***. В шаблонах этой группы используется вспомогательный шаблон с именем `Term`, описывающий возможные морфосинтаксические образцы определяемого термина, и шаблон `Defin`, задающий структуру определяющей термин конструкции.

Таблица 1. Лексико-синтаксические шаблоны

№	Группы шаблонов	Примеры шаблонов	Примеры терминов и их употреблений
1	Морфо-синтаксические образцы терминов	`N1 (N1)`	*вимп*
		`A1 N1 <A1=N1> (N1)`	*опорная точка*
		`N1 N2<c=gen> (N1)`	*период упреждения*
		`N1 A2 N2<c=gen> <A2=N2> (N1)`	*технология двойной накачки*
2	Контексты определения авторских терминов	`Defin<c=acc> "будем" "называть" Term<c=ins> #Term<c=nom>`	*Такие операции будем называть понятийными операциями*
		`"под" Term<c=ins> "понимается" Defin<c=nom> #Term<c=nom>`	*Под прерыванием понимается сигнал…*
3	Контексты введения синонимов терминов	`Term1 "("Term2")" <Term1.c=Term2.c> #Term1<c=nom>, Term2<c=nom>`	*взаимодействующих компонентов (подсистем)*
		`Term1 "," "или" Term2 <Term1.c=Term2.c> #Term1<c=nom>, Term2<c=nom>`	*разрядностью, или длиной слова*
4	Словарные термины	`N1<вектор> [N2<намагниченности,c=gen>| N2<состояния,c=gen>|"Умова"]`	*вектор, вектор намагниченности, вектор состояния, вектор Умова*
		`A1<битовый> {N2<массив>| N2<образ>}<1,1> <A1=N2>`	*битовый массив, битовый образ*
5	Лексико-синтаксические варианты	`N1 N2<c=gen> #N1, N1 N4<c=gen> <Syn(N2,N4)>, N3 N2<c=gen> <Syn(N1,N3)>, A1 N1 <A1.st=N2.st>`	*вывод данных — вывод* (`N1`), *вывод информации* (`N1 N4`)
			шина адреса — шина (`N1`), *адресная шина* (`A1 N1`)
6	Соединения терминов	`"как" A1 "," "так" "и" A2 N1 <A1=A2=N1> #A1 N1, A2 N1`	*как тонкий, так и толстый клиент — тонкий клиент, толстый клиент*
		`N1 N2<c=gen> "," N3<c=gen> {"и"|"или"} N4<c=gen> #N1 N2<c=gen>,N1 N3<c=gen>, N1 N4<c=gen>`	*шинам адреса, данных и управления — шина адреса, шина данных, шина управления*
		`A1 A2 N1 <A1=A2=N1> #A1 N1, A2 N1`	*удаленный банковский терминал — банковский терминал, удаленный терминал*
		`N1 A2 N2<c=gen> <A2=N2> #N1 N2, A2 N2`	*разрядность внутренних регистров — разрядность регистра, внутренний регистр*

Заметим, что все шаблоны второй группы включают в свой состав элемент `#Term <c=nom>` — выделяемую конструкцию. В общем случае *выделяемая конструкция* (записывается после знака #) определяет, какая часть распознанной конструкции должна быть из нее выделена и с какими морфологическими характеристиками. В данном случае задается выделение распознанного авторского термина, причем в нормализованной форме — в именительном падеже (для рассмотренного примера — *вимп*).

Третья группа шаблонов строилась аналогично второй — в нее входят шаблоны типичных контекстов, в которых вводятся синонимы терминов, например: *…проектирование прикладного ПО (приложений)*. В шаблонах задается выделение пары синонимичных терминов (*прикладное ПО* и *приложения*).

Язык лексико-синтаксических шаблонов оказался удобным и для записи входов терминологических словарей (четвертая группа шаблонов). Каждый шаблон позволяет описать в общем случае несколько *словарных* терминов, имеющих общее начало.

Последние две группы шаблонов описывают в общем виде текстовые варианты терминов: *лексико-синтаксические варианты* одного термина и *соединения* нескольких терминологических словосочетаний (в Табл. 1 в соответствующих примерах шаблонов обоих групп для краткости опущены условия нормализации).

Шаблоны пятой группы по сути формализуют правила образования вариантов термина: каждый шаблон фиксирует один из морфосинтаксических образцов термина и задает (после знака #) вы-

деляемые конструкции — возможные текстовые варианты термина (тоже в виде морфосинтаксических образцов). Например, для терминологических словосочетаний со структурой `N1 N2<c=gen>` (см. строку 5 Табл. 1) учтены следующие случаи:

1) Вставка или отбрасывание слова — вариант `N1` (*ввод данных — ввод*).
2) Замена одного слова на синоним в данной ПО — варианты `N1 N4<c=gen>` (*вывод информации — вывод данных*) и `N3 N2<c=gen>` (*метка адреса — маркер адреса*), при этом предполагается выполнение соответствующих словарных условий синонимии.
3) Замена слова на однокоренное другой части речи с одновременным изменением синтаксических связей словосочетания — вариант `A1 N1 <A1=N1>` (*шина адреса — адресная шина*), условие равенства корней слов записывается как `A1.st=N2.st`.

Аналогично были формализованы правила образования типичных *соединений* в тексте нескольких терминологических словосочетаний, при которых один или несколько терминов разрываются и/или усекаются. Среди соединений терминов мы различаем:

- соединения с помощью сочинительных союзов (*шина адреса, данных и управления — шина адреса*, *шина данных*, *шина управления)* и
- бессоюзные соединения (*разрядность внутренних регистров — разрядность регистров*, *внутренние регистры*).

Каждый шаблон соединения задает выделение всех входящих в него терминов.

3. Процедуры распознавания и их тестирование

Для каждой из описанных групп лексико-синтаксических шаблонов была разработана процедура автоматического распознавания в научно-техническом тексте соответствующих терминоупотреблений. Эти процедуры позволяют выявлять соответственно термины-кандидаты, авторские термины, термины-синонимы, словарные термины, лексико-синтаксические варианты и соединения терминов. Дополнительно, процедуры подсчитывают частоту употребления каждого из распознанных ими терминов. Результат распознавания терминов по их морфосинтаксическому образцу назван нами *терминами-кандидатами*, чтобы подчеркнуть, что в их числе с достаточно большой вероятностью могут оказаться общенаучные словосочетания вида *решение задачи*, *применение последнего правила*, не являющиеся терминами.

На вход каждой процедуре, за исключением процедуры выявления лексико-синтаксических вариантов, поступает анализируемый текст и соответствующая группа шаблонов. На первом этапе работы процедуры выполняется поиск всех фрагментов текста, представляющих собой искомые терминоупотребления, а на втором — подсчитывается частота употребления каждого выявленного в них термина. Разделение этих двух этапов необходимо для корректного подсчета частоты употребления в случаях полного вложения одного термина в другой (*адрес — логический адрес*), поскольку частота должна быть определена без учета таких вложений. **Таким образом, на выходе** указанных процедур получается список фрагментов-терминоупотреблений и список выделенных терминов с частотой их употребления в тексте.

В процедурах, опирающихся на шаблоны с выделяемыми конструкциями (шаблоны авторских терминов, синонимов и соединений), этап выявления происходит в два приема. Сначала в тексте ищутся все фрагменты, соответствующие шаблонам и из них выделяются термины, точнее шаблоны, их описывающие, и эти шаблоны используются затем для поиска всех употреблений этих терминов. В частности, при работе процедуры выявления авторских терминов сначала ищутся контексты определения терминов, из них выделяются авторские термины, вхождения которых потом ищутся в тексте.

По иному организована процедура выявления лексико-синтаксических вариантов терминов. Н**а вход** ей поступают:

- шаблоны правил образования лексико-синтаксических вариантов;
- термины, для которых необходимо найти их варианты;
- слова и словосочетания, среди которых необходимо искать эти варианты.

На выходе процедуры получаются группы эквивалентных вариантов; каждая группа объединяет слова и словосочетания, соответствующие (предположительно) одному и тому же понятию ПО.

Все разработанные процедуры были по отдельности протестированы на научно-технических текстах из разных областей физики и информатики (общего объема 700 Кбайт). При этом для выявления словарных терминов использовались словари по физике (более 3 тыс. терминов) и по информатике (более 4 тыс. терминов). Для выявления лексико-синтаксических вариантов был создан рабочий словарь терминологических синонимов в этих ПО.

Примерно для трети текстов результаты работы процедур были сравнены со списками терминов, выявленными в текстах экспертами, при этом оценивались полнота и точность как выделения самих терминов, так и их употреблений в тексте — см. Табл. 2. Для процедур распознавания синонимов и соединений замерялась полнота и точность выделения терминов, встретившихся в рамках распознаваемых ими конструкций.

Таблица 2. Полнота и точность процедур

Процедура	Выделение терминов		Выделение терминоупотреблений	
	Полнота	Точность	Полнота	Точность
Термины-кандидаты	58 %	24 %	54 %	25 %
Авторские термины	67 %	89 %	70 %	97 %
Синонимы	57 %	22 %	–	–
Словарные термины	85 %	94 %	87 %	95 %
Соединения	71 %	30 %	–	–

Наихудшие результаты (что было вполне прогнозируемо) дала процедура выявления терминов-кандидатов, опирающаяся на минимум лингвистической информации. Она давала много «шума» (*крупный размер, аналогичный результат*), в тоже время не были *распознаны сочетания, чья морфосинтаксическая структура не учтена в наборе шаблонов (например, термины вида индекс iCOMP и обратная связь по релевантности)*.

Для **словарных терминов** нераспознанными оказались преимущественно их терминоупотребления внутри соединений, *а некоторые распознанные термины* оказались частью несловарных терминов (например, термин *ряд* — частью общенаучных выражений: *в ряде случаев*, *за рядом исключений)*.

Для **авторских терминов** и синонимов основная потеря полноты возникла опять же из-за ограниченности морфосинтаксических образцов терминов, но также и по причине неучета некоторых конструкций определения терминов и их синонимов (к примеру, *Регистр представляет собой совокупность бистабильных устройств*). Подобные конструкции рассматривались нами при построении набора шаблонов, но не были включены в него, поскольку часто имеют смысл, отличный от определения. Неточность же выявления терминов касалась тех случаев, когда в характерном контексте определения термина уточнялось значение словарных терминов.

Что касается процедур выявления соединений, то причины невысоких значений полноты и точности такие же, как и для группы терминов-кандидатов.

4. Стратегия совместного применения процедур

В целом, проведенный анализ результатов работы процедур показал, во-первых, что полноту распознавания можно повысить, расширяя набор шаблонов, но при этом часто страдает точность. Во-вторых, возможный учет процедурами результатов работы других процедур в ряде случаев может повысить полноту распознавания: так, в случае учета выявленных из соединений разрывных терминов наблюдался прирост полноты в среднем на 12 %. В-третьих, использование результатов других процедур позволяет более точно определять частоту употребления каждого конкретного термина в тексте.

Поскольку простое объединение результатов работы процедур (вырабатываемых ими списков терминов) повышает полноту выявления терминов, достаточно сильно снижая точность, были изучены другие способы объединения. Это позволило сформулировать соответствующую стратегию совместного применения процедур, согласно которой:

1) К заданному тексту применяются все процедуры распознавания, за исключением процедуры поиска лексико-синтаксических вариантов.
2) Распознанные словарные и авторские термины, объединяются и включаются в формируемое множество выделенных терминов, причем в случае полных вложений предпочтение отдается более длинным терминам.
3) Из выявленных терминов-кандидатов в полученное множество включаются только те, в состав которых входят словарные или авторские термины, при этом последние исключаются из формируемого множества.
4) Из найденных пар терминологических синонимов берутся только пары, один член которых уже входит в формируемое множество, и в него добавляется второй член пары.
5) Термины из выявленных соединений добавляются во множество, только если среди них есть разрывный словарный термин.
6) Для сформированного множества терминов применяется процедура поиска их лексико-синтаксических вариантов из числа оставшихся терминов-кандидатов, и полученные варианты добавляются к формируемому множеству.
7) Во множество дополнительно добавляются термины из тех соединений, в которые входят текстовые варианты, найденные на предыдущем этапе.
8) Из оставшихся терминов-кандидатов в формируемое множество добавляются те, частота употребления которых выше заранее установленного порога (например, равного среднему квадратичному частот терминов-кандидатов).

Поскольку показатели полноты и точности распознавания взаимосвязаны (как правило, увеличение одного показателя приводит к уменьшению другого) для оценки результатов рассмотренной стратегии использовалась F-мера — комбинированный

показатель, вычисляемый как гармоническое среднее полноты и точности. Оказалось, что в большинстве случаев применение стратегии дает ощутимое повышение F-меры по сравнению с простым объединением результатов работы процедур распознавания. К примеру, для нижеследующего текста был зафиксирован прирост F-меры выявления терминов на 19,8 %, а F-меры выявления терминоупотреблений — на 15,5 % (все выявленные терминоупотребления подчеркнуты: двойной линией подчеркнуты слова, являющиеся частями нескольких терминов, пунктирной линией — слова, являющиеся частями разрывных терминов):

Микропроцессор, как правило, представляет собой сверхбольшую интегральную схему, реализованную в едином полупроводниковом кристалле и способную выполнять функции центрального процессора. С внешними устройствами микропроцессор может «общаться» благодаря шинам адреса, данных и управления, выведенным на специальные контакты корпуса микросхемы. Стоит отметить, что разрядность внутренних регистров микропроцессора может не совпадать с количеством внешних выводов для линий данных. Иначе говоря, микропроцессор с 32-разрядными регистрами может иметь, например, только 16 линий внешних данных.

Любое внешнее устройство, совершающее по отношению к микропроцессору операции ввода-вывода, можно назвать периферийным.

Порт — это некая схема сопряжения, обычно включающая в себя один или несколько регистров ввода-вывода и позволяющая подключить, например, периферийное устройство к внешним шинам микропроцессора. Практически каждая микросхема использует для различных целей несколько портов ввода-вывода.

В приведенном тексте несколько терминов остались нераспознанными (например, *контакты корпуса* — по причине отсутствия в словаре по информатике), в то же время выявлено как термин общенаучное словосочетание *различные цели*. Дальнейшая настройка стратегии требует проведения дополнительных экспериментов.

5. Заключение

Разработаны процедуры выявления терминоупотреблений в заданном научно-техническом тексте на основе набора лексико-синтаксических шаблонов, полученных в ходе формализации различных случаев употребления в текстах терминологических слов и словосочетаний. На основе анализа результатов их раздельного тестирования предложена стратегия совместного применения этих процедур, позволяющая повысить F-меру полноты и точности распознавания. Для научной прозы дальнейшие резервы повышения полноты связаны с учетом дополнительных видов текстовых вариантов, а точности — с учетом взаимного расположения вариантов в тексте и словаря общенаучных выражений.

Поскольку используемый набор шаблонов может быть изменен без перепрограммирования процедур распознавания, это дает возможность настройки процедур на обработку текстов разных ПО с учетом присущих им особенностей терминоупотребления.

Литература

1. *Jacquemin C., Bourigault D.* Term extraction and automatic indexing // Mitkov R. (ed.): Handbook of Computational Linguistics. Oxford University Press, 2003. P. 599–615.
2. *Добров Б. В., Лукашевич Н. В., Сыромятников С. В.* Формирование базы терминологических словосочетаний по текстам предметной области // Труды пятой всероссийской научной конференции «Электронные библиотеки: Перспективные методы и технологии, электронные коллекции». 2003. С. 201–210.
3. *Nenadic G., Ananiadou S., McNaught J.* Enhancing Automatic Term Recognition through Variation // Proceedings of 20th Int. Conference on Computational Linguistics COLING'04. 2004. P. 604–610.
4. *Большакова Е. И., Баева Н. В., Бордаченкова Е. А., Васильева Н. Э., Морозов С. С.* Лексико-синтаксические шаблоны в задачах автоматической обработки текстов // Компьютерная лингвистика и интеллектуальные технологии: Труды Международной конференции Диалог '2007. М.: Изд-во РГГУ, 2007. С. 70–75.
5. *Носков А. А.* Метод выделения в тексте конструкций по их лексико-синтаксическим шаблонам // Сборник статей молодых ученых факультета ВМиК МГУ. М.: Издательский отдел фак-та ВМиК МГУ им. М. В. Ломоносова; МАКС Пресс, 2009. Выпуск 6. С. 136–145.
6. *Большакова Е. И., Васильева Н. Э.* Терминологическая вариантность и ее учет при автоматической обработке текстов // Одиннадцатая национальная конференция по искусственному интеллекту с международным участием: Труды конференции. М.: ЛЕНАНД, 2008. Т. 2. С. 174–182.

О месте видовых троек в аспектуальной системе русского языка

Aspectual Triplets in Contemporary Russian Aspectual System

Зализняк Анна А. (anna-zalizniak@mtu-net.ru)
Институт языкознания РАН

Микаэлян И. Л. (irina-mikaelian@yandex.ru)
The Pennsylvania State University, USA

Видовые тройки представляют собой регулярное явление, определяющее облик русской аспектуальной системы, поскольку они возникают в результате действия того же механизма, который обеспечивает функционирование видовых пар. Существование видовых троек не отменяет понятия видовой пары, конституирующей для русского вида как грамматической категории.

Наше исследование основано на концепции, восходящей в идеям Ю. С. Маслова [Маслов1984] и изложенной, в частности, в работах [Зализняк, Шмелев 2000; 2001, 2004, Mikaelian, Shmelev, Zalizniak 2007; Зализняк, Микаэлян, Шмелев 2010]. Мы исходим из того, что видовая пара в русском языке — это функциональная корреляция, определяемая отношением субституции, т. е. замены глагола сов. вида на некоторый глагол несов. вида в тех контекстах, где, по правилам русской грамматики, несов. вид не может быть употреблен. Этот принцип субституции основан на следующей особенности русской аспектуальной системы. Глаголы сов. вида всегда обозначают **события**, а глаголы несов. вида обычно обозначают **процессы** или **состояния**, но при определенных условиях могут обозначать также и **события**. Такими условиями, в частности, являются: повествование в наст. историческом и контекст многократности. Это «контексты обязательной имперфективации», или «контексты Маслова». Так, на месте глаголов сов. вида *пришел, увидел, победил* при повествовании в наст. историческом окажутся глаголы несов. вида *приходит, видит, побеждает*. Эти глаголы являются имперфективными коррелятами к тем глаголам сов. вида, «на месте» которых они появились, и образуют с ними видовые пары. В данном случае для каждого глагола сов. вида такой имперфективный коррелят один: *приходить* для *прийти, видеть* для *увидеть, побеждать* для *победить*. Между тем в русском языке существуют такие глаголы сов. вида, для которых сформулированному выше требованию удовлетворяют два глагола несов. вида. Так, например, глаголу сов. вида *скомкать* в предложении *Он прочел записку, скомкал ее и выбросил в корзину* при пересказе в наст. историческом можно сопоставить любой из двух глаголов: *комкать* или *скомкивать*. В таких случаях говорят о ***видовых тройках***.

Видовым тройкам посвящена обширная литература (см. в частности [Апресян 1995; Ясаи 2001; Петрухина 2001; Храковский 2005; Гиро-Вебер, Микаэлян 2006; Зализняк, Микаэлян 2007; Guiraud-Weber 2004; Janda 2007;]). Наше обращение к проблеме видовых троек вызвано двумя причинами. С одной стороны, в последние годы приобретает популярность идея, что видовых пар не существует, а есть только «кластеры» — пучки деривационно связанных глаголов разного вида (в том числе, видовые тройки), с чем мы не можем согласиться. С другой стороны, мы видим нашу задачу в том, чтобы выяснить картину реального функционирования в современном русском языке глаголов несов. вида, входящих в видовые тройки — что стало возможно именно сейчас, благодаря наличию электронных корпусов и поисковых систем в Интернете[1].

Мы хотим показать следующее.

Во-первых, видовые тройки представляют собой не периферийное, а в высшей степени ре-

[1] Все примеры словоупотребления, которые мы приводим без ссылок, получены из Интернета при помощи поисковой системы Яндекс.

гулярное явление, определяющее облик русской аспектуальной системы, поскольку они возникают в результате действия того же механизма, который обеспечивает наличие имперфективного коррелята почти для любого глагола сов. вида — ср. формы типа *рухать*, *нагрядывать*, *уцелевать* и т. п., широко употребительные не только в разговорной речи (о чем можно судить по данным Интернета), но также и в художественной литературе (ср. [Гловинская 2008]). Действительно, наличие имперфективного коррелята у любого глагола сов. вида — это чисто системное требование, и для его реализации имеется универсальное морфологическое средство: суффикс *-ыва-/-ива-*: так возникают глаголы типа *заблуживаться*, *встрепеныватьсяся*; они появляются, уходят и снова появляются, вспомним гоголевские формы типа *помарщивается* или *посторанивают-ся*. В Интернете можно встретить формы типа *нарисовывать* или *скушивать*, которые обнаруживаются у Даля, и т. д. Это важнейший, очень мощный механизм морфологической имперфективации, который воздействует на любой глагол сов. вида[2]. Если он применяется к глаголу сов. вида, не имеющему бесприставочного имперфективного коррелята, то возникает суффиксальная видовая пара; она может быть словарной (ср. *опоздать* → *опаздывать*) или же окказиональной (ср. *уцелеть* → *уцелевать*, *нагрянуть* → *нагрядывать*). Если же глагол сов. вида уже имеет бесприставочный имперфективный коррелят, тот же механизм порождает тройку: например, *намазать* → *намазывать*, при уже имеющемся *мазать*. При этом граница между этими двумя случаями является нежесткой; тем самым видовые тройки и видовые пары образуют некую единую систему (см. ниже Табл.1).

Во-вторых, мы хотим показать, что существование видовых троек никак не компрометирует понятие видовой корреляции как бинарного отношения, возникающего между глаголом совершенного и несовершенного вида в контекстах обязательной имперфективации. Однако бинарность видовой корреляции как отношения субституции затемняется тем, что здесь в игру вступают еще по крайней мере два фактора:

1) **префиксальное словообразование**, в результате которого возникает новый глагол; его собственно семантическое расхождение с исходным глаголом может быть большим или меньшим, но он всегда сов. вида (что порождает сходство и, как следствие, смешение с собственно аспектуальной корреляцией);

2) **лексическая синонимия**, связывающая между собой как глаголы сов. вида с разной приставкой (ср. *примерить* и *померить*, *поменять* и *обменять*), так и глаголы несов. вида, один из которых является первичным, а другой — вторичным имперфективом, например: *капать/закапывать* <капли в нос>, *множиться/умножаться*, *гибнуть/погибать*, *клеить/склеивать* <коробочку>. Вообще в процессе порождения высказывания говорящий довольно часто сталкивается с проблемой выбора из ряда синонимов; эта проблема может возникать, в том числе, при выборе глагола несов. вида в контекстах обязательной имперфективации. Соответственно, говорящий оказывается вынужден выбирать между *мерить* и *примерять*, *мазать* и *намазывать* и т. п. — но это уже проблема не аспектологии, а лексической семантики.

Среди факторов, затемняющих бинарную природу видовой оппозиции, можно назвать еще некоторые чисто морфологические обстоятельства, блокирующие образование имперфектива или форм деепричастия, и в этом случае может использоваться, в качестве second best, другой, близкий по смыслу глагол, например, можно сказать только *погибая*: форма деепричастия от глагола *гибнуть* в русском языке не образуется.

Эти два основных фактора — префиксальное словообразование и лексическая синонимия — могут и должны быть отделены от собственно аспектуальной корреляции. Отказ от этого разграничения и, соответственно, от категории видовой пары — в пользу «кластеров», в которых все эти три параметра (вид, префиксация и синонимия) соединены, как нам представляется, нисколько нас не продвигает в познании природы вещей, а наоборот, является шагом назад. Трудности же, которые возникают при попытке разграничения упомянутых выше параметров, как мы попытаемся показать, могут быть преодолены путем использования аппарата градуальных признаков.

Итак, пусть имеется бесприставочный глагол несов. вида (НСВ1 — например, *мазать*), образованный от него префиксацией глагол сов. вида (СВ — например, *намазать*), и образованный от последнего путем имперфективизации вторичный имперфектив (НСВ2: *намазывать*). Таким образом возникают ***морфологические биимперфективные тройки***, которые и будут объектом дальнейшего рассмотрения[3]. Если морфологическая тройка оказывается видовой, то мы будем записывать ее (вопреки имеющейся традиции, в соответствии с логикой аспектуальной корреляции, а не морфологической деривации) в форме «СВ — НСВ1/НСВ2», ср.: *намазать — мазать/намазывать*.

---

[2] Об имеющихся ограничениях действия этого механизма см. [Зализняк, Шмелев 2000: 84–85, Зализняк, Микаэлян, Шмелев 2010].

[3] Тем самым мы исключаем, с одной стороны, тройки с морфологической вариативностью внутри НСВ2 (*простудиться — простужаться/простуживаться*, ср. 1-й аспектуальный тип по [Апресян 1995]) и, с другой стороны, биперфективные тройки типа *помыть/вымыть — мыть*, представляющие собой отдельный феномен.

Очевидно, что не все такие тройки глаголов являются «видовыми», так как для видовой тройки требуется, чтобы оба глагола несов. вида **претендовали на роль видового коррелята** к глаголу сов. вида. Дело в том, что вопрос, может ли данный глагол несов. вида замещать данный глагол сов. вида в контекстах Маслова допускает не только ответы «да» или «нет»: для значительного количества русских глаголов ответ будет «в крайнем случае (в каких-то случаях, иногда и т. п.) может»; кроме того, разные носители языка могут дать различные ответы. Именно это обстоятельство отражено в формулировке «претендует на роль» имперфективного коррелята (см. об этом ниже).

В класс видовых троек не входят, хотя и непосредственно примыкают к нему, по крайней мере два типа морфологических троек:

(22) *писать* → *переписать* → *переписывать*;
писать (отдельный глагол)
переписать — *переписывать* (суффиксальная видовая пара)

(23) *шить* → *сшить* → *сшивать*
$сшить_1$ <платье> — *шить* (депрефиксальная видовая пара)
$сшить_2$ <два куска материи> — *сшивать* (суффиксальная видовая пара)

Первый из них не входит в класс троек, поскольку глагол НСВ1 *писать* не претендует на роль видового коррелята к *переписать*, т. е. не может обозначать того же события, которое обозначается глаголом СВ *переписать*. Тем самым здесь мы имеем не тройку, а в чистом виде суффиксальную пару плюс «отдельный» глагол НСВ1. Поэтому из числа видовых троек будут исключены такие морфологические тройки, в которых вторичный имерфектив образует регулярную пару с глаголом СВ, а мотивирующий глагол НСВ на такую роль не претендует. При этом мотивирующий глагол НСВ1 может входить в другую пару — депрефиксальную — как в случае с *писать* (который образует пару с глаголом СВ *написать*) и множеством других глаголов.

На первый взгляд, для исключения такого типа морфологических троек из рассмотрения было бы достаточно поставить семантическое условие: действительно, *переписать,* или *надстроить*, или *развязать* образованы присоединением дополнительной полнозначной морфемы (в отличие от *написать*, *построить*, *связать*, и т.д, где приставка десемантизирована). Тем самым это совершенно «другие глаголы», чем *писать, строить* и *вязать*. Так о какой тройке здесь может идти речь? Однако собственно семантический критерий («глагол с другим значением или с тем же самым») здесь оказывается неприменим по двум причинам. Во-первых, не существует жесткой границы между «полнозначной» и «десемантизированной» приставкой. Во-вторых, существует механизм, который позволяет бесприставочному глаголу НСВ1, в каких-то условиях, «брать на себя», в том числе достаточно специфическое, значение приставочного глагола НСВ2 — это случаи типа *писать на пленку* (вместо *записывать*) или *варить рельсы* (вместо *сваривать*), о которых пойдет речь ниже (см. примеры (6), (7) и (8) и комментарии к ним).

За пределами класса видовых троек окажутся также морфологические тройки типа *искать — отыскать — отыскивать*; *искать — разыскать — разыскивать*, *знать — узнать — узнавать*, где глагол НСВ1 неспособен обозначать никакое событие (и поэтому не входит ни в видовую тройку, ни в видовую пару).

Второй тип не относится к классу видовых троек, поскольку такая морфологическая тройка распадается на две пары, в соответствии с двумя значениями глагола СВ, ср. (2). Различие между двумя значениями глагола *сшить* определяется значением приставки *с-*: в случае *сшить* <платье> она имеет собственно результативное («чистовидовое») значение и, соответственно, глагол имеет депрефиксальный коррелят *шить*, в случае *сшить* <два куска материи> приставка имеет самостоятельное значение «соединения», т. е. это новый глагол с полнозначной приставкой, соответственно, используется суффиксальный коррелят — *сшивать*. Таким образом морфологическая тройка распадается на две пары.

Оба эти случая (тип *переписать* и тип *сшить*) хотелось бы просто исключить из рассмотрения. Этого, однако, сделать нельзя по следующей причине. Действительно, приведенные в примерах (1) и (2) морфологические тройки представляют собой «чистые» случаи: они никак не могут быть отнесены к видовым. Однако такие случаи довольно редки: уже глаголы *записать* и *склеить*, близкие по форме и по значению, представляют более «смазанную» картину. Так, нельзя сказать **она каждый месяц сшивает себе новое платье* (надо сказать *шьет*). Но сказать, например, *она каждый месяц склеивает ребенку новую картонную коробочку для игрушек* — можно, наряду с *клеит*, т. е. здесь имеется конкуренция (по крайней мере, в данном значении), и тем самым, это собственно видовая тройка. Таким образом, тройка *склеить — клеить/склеивать*, оказывается устроена иначе, чем *сшить — шить/сшивать*.

Если мы возьмем глагол СВ *записать* <исполнение музыкального произведения на пленку>, то окажется, что наряду со стандартным имперфективным коррелятом НСВ2 *записывать* в профессиональной речи употребляется глагол НСВ1 *писать* (в том числе, и с событийном значении, т. е. в контекстах Маслова). И при этом такая «подмена» (употребление, в профессиональной или разговорно-фамильярной и т. д. речи НСВ1 вместо стандартного литературного НСВ2) — явление

весьма регулярное, игнорировать которое лингвист не вправе, ср. *рвать зубы, рвать мосты, драть нос, жечь документы, крепить балку, копать картошку, варить трубы* и т. п. Возникает ли в таких случаях видовая тройка (ср. *вырвать* <зуб> — *рвать/вырывать*), зависит как минимум от двух факторов: имеет ли такой НСВ1 событийное значение (иногда имеет: *рвут мосты и идут дальше*) и имеется ли дополнительное распределение НСВ1 и НСВ2 по функциональным стилям.

Тем самым, чтобы определить, является ли некоторая морфологическая тройка видовой и если да, то в какой мере, необходимо исследование, включающее поиск примеров в корпусах и их оценку с точки зрения допустимости.

Так, к классу «невидовых троек» типа 2 относятся, например:

(24) а. *поставить* <чайник на плиту> — *ставить*
<сырье на завод> — *поставлять*
б. выпороть <ученика> — *пороть*
<нитку> — *выпарывать*
в. замочить <кого-то> [= убить] — *мочить*
<белье> — *замачивать*

Здесь везде имеется полное распадение на две пары, и никакой тройки не возникает. Но, например, глагол *сварить* устроен иначе. В литературном языке от относится к тому же «чистому» типу, с четким дополнительным распределением, которое исключает возникновение тройки:

(25) *сварить* <суп> — *варить*
<рельсы > — *сваривать*

Однако в профессиональной речи сварщиков употребляется также имперфектив *варить*. Тем самым глагол *сварить* в значении ‘соединить при помощи сварки’ должен быть признан образующим видовую тройку:

(26) *сварить* <рельсы> — *варить/сваривать*

Так, поисковая система Яндекс нашла для сочетания *сваривать рельсы* 168 тысяч употреблений, для *варить рельсы* — 207 тысяч. Ср. примеры (6), (7) и (8), в которых оба глагола (*сваривать* и *варить*) имеют событийное значение:

(27) *Теперь красноярские железнодорожники будут* ***сваривать*** *рельсы за считанные секунды.*

(28) *…выкраивали прямые кусочки по 60 — 120 сантиметров, из которых потом* ***варили*** *рельсы десятиметровой длины, годные для укладки;*

(29) *… приходит газовщик, крутит что-то в плите, заявляет, что нужен сварщик — переварить трубу. Рекомендует своего знакомого. Приходит сварщик,* ***варит*** *трубу, теперь нужен штукатур — чтоб убрал следы на потолке.*

В области частичного распадения потенциальной тройки на две пары, в соответствии с разными значениями или типами употребления, представлены следующие возможности:

- **в одном значении — НСВ1 /НСВ2, в другом — НСВ2:**
схватить <кого-то за руку> — *хватать/схватывать*
<бревна железной скобой> — *схватывать*,

Для глагола *схватить* в центральном значении (*схватить кого-то за руку*) имперфективным коррелятом является *хватать* (но *схватывать* тоже встречается), а в значении ‘соединить, скрепить’ (*схватить бревна железной скобой*), так же, как и в ряде переносных (*схватить насморк*, и др.) употребляется лишь вторичный имперфектив *схватывать*.

- **в одном значении — НСВ1 /НСВ2, в другом — НСВ1:**
разделить <яблоко на три части, студентов на группы> — *делить/разделять*; <наследство; обязанности; пятнадцать на три> — *делить*;
сварить <рельсы > — *варить/сваривать;* <суп> — *варить*.

Видовые тройки (т. е. морфологические тройки, в которых НСВ1 и НСВ2 претендуют на статус коррелята к соответствующему глаголу СВ) представляют собой неоднородный класс. Для них существенно противопоставление по следующим признакам (все они градуальны):

1. Степень синонимичности (взаимозаменимости) НСВ1 и НСВ2. Условно здесь можно выделить два типа случаев: (а) случаи полной синонимии, т. е. свободного варьирования (очень редкие) и (б) случаи частичной синонимии (которых большинство).

а) *гибнуть* и *погибать*, *публиковать* и *опубликовывать*. По-видимому, сюда относятся также *множиться* и *умножаться*, *слабеть* и *ослабевать* и другие глаголы с приставкой *о-*, означающие приобретение свойства.

б) *учить/выучивать*, *мазать/намазывать*; *лезть/залезать* <на дерево> и т. п. (см. вторую и третью строку в таблице 1)[4].

2. Степень конвенционализации событийного значения у НСВ1. Способность каждого из глаголов НСВ1 и НСВ2, хотя бы в каких-то условиях, иметь событийное значение (т. е. заменять глагол СВ в контекстах Маслова) является неотъемлемым для понятия видовой тройки. Однако разные глаголы НСВ об-

[4] Проведенное нами обследование глаголов несов. вида Ожегова-Шведовой позволило выделить около 160 таких троек.

ладают этой способностью в разной степени. Вообще надо сказать, что событийное значение у глаголов несов. вида, за исключением тривиальных видовых коррелятов типа *достигать, съедать* или *приходить* (т. е. тех немногих глаголов несов. вида, которые не имеют никакого «собственного», несобытийного значения) всегда контекстно обусловлено. Иными словами, событийное значение для глагола несов. вида — это всегда лишь способность, потенция — но при этом некоторые глаголы несов. вида даже такой потенциальной событийности не содержат. Тем самым, наиболее существенная граница проходит не между глаголами, «имеющими» и «не имеющими» событийное значение, а между такими как *ловить*, *решать*, *думать*, *чувствовать* и даже *лечить*, *учить*, *есть и пить*, которые, при определенных условиях, могут обозначать событие, и глаголами типа *знать*, *любить* или *искать,* которые ни при каких обстоятельствах событийного значения выразить не смогут. Анализ большого массива глаголов и текстов (мы брали подряд весь словник словаря Ожегова-Шведовой и проверяли потенциальные тройки при помощи поиска в НКРЯ и в Яндексе) показывает, что способность иметь событийное значение представляет собой шкалу, где между глаголами, имеющими и не имеющими событийное значение, существует целый спектр промежуточных случаев. Приведем несколько примеров контекстно обусловленного событийного употребления глаголов НСВ1[5]:

(30) ***ем*** *в день одно яблоко и* ***пью*** *стакан воды*;

(31) *мануальный терапевт Валентина Васильевна* ***лечит*** *за один сеанс;*

(32) *отвердитель все же* ***густеет за*** *несколько часов на воздухе.*

3. Существенной характеристикой видовых троек является также наличие/отсутствие у глагола НВС2 **процессного** значения. Существует мнение, что в тройках глаголы НСВ1 и НСВ2 распределены по значениям, так что НСВ1 имеет лишь процессное значение, а НСВ2 — лишь событийное. Сразу скажем, что такие тройки, если бы они существовали, вообще не являлись бы видовыми в том смысле, как они определены в данной статье. Однако, как было сказано выше, многие глаголы НСВ1, у которых традиционно выделяется только процессное значение, могут, в определенных условиях, обозначать событие (см. примеры (9)-(11)). С другой стороны, как показано в [Петрухина 2000], среди глаголов НСВ2 очень немного таких (как *съедать*) которые не имеют процессного значения. Тем самым, оказывается, что видовые тройки — гораздо более системное явление в русском языке, чем это принято считать.

4. Вхождение НСВ1 в другую видовую пару или видовую тройку (так, глагол *мерить* входит как в тройку *примерить — мерить/примерять*, так и в пару *померить — мерить*). В таких случаях естественно считать глагол *примерять* собственно видовым коррелятом *примерить*, а глагол *мерить,* соответственно, коррелятом глагола *померить*. Такое решение будет системно последовательным, но одновременно и упрощенным.

5. Степень употребительности (приемлемости) вторичного имперфектива: шкала от единственно возможной формы (*открывать*, *перепечатывать*), через контекстно ограниченные формы (ср. *сламывать* <человека>, ??<ветку>, *<дом>), к виртуально существующим формам (типа *нарисовывать*), встречающимся в языке Интернета и в детской речи [Цейтлин 2009: 313]) и далее, к практически несуществующим (но все же тоже встречающимся, ср. *постраивать* <схему>). Оценка приемлемости НСВ2 очень сильно варьирует в зависимости от: 1) времени (меняется с течением времени); 2) социальной, профессиональной, возрастной и пр. группы говорящих; 3) идиолекта. Те глаголы, которые кажутся «искусственными», «тяжеловесными», «неуклюжими» и т. д. одним говорящим (исследователям), другим представляются совершенно обычными словами русского языка. Не говоря уже о том, что даже самые причудливые имперфективы в огромном количестве встречаются в Интернете (такие как *заблуживаться*, *заподазривать*, *встрепеныватьс*я…., см. материал на эту тему в [Mikaelian, Shmelev, Zalizniak, 2007]).

Полное описание русской аспектуальной системы должно включать характеристику каждой видовой тройки по всем этим параметрам, так как именно совокупность значений данных параметров определяет функциональный тип видовой тройки и его место в системе русского глагола.

Место видовых троек в аспектуальной системе русского языка может быть продемонстрировано при помощи таблицы 1.

Итак, видовые тройки — неотъемлемая составляющая русской аспектуальной системы. Соотношение понятий видовой тройки и видовой пары может быть охарактеризовано следующим образом. Видовая пара — это абстрактное отношение субституции в контекстах обязательной имперфективации, формирующее русский вид как грамматическую категорию. Тот факт, что многие русские приставочные глаголы сов. вида вступают в такую корреляцию с двумя разными глаголами несов. вида означает только, что в этом участке системы, помимо собственно аспектуальной корреляции, в игре участвуют также парадигматические отношения лексической синонимии. Видовая тройка — результирующая этих двух факторов.

[5] Имеющийся у нас обширный материал, иллюстрирующий подобное употребление НСВ1, мы здесь не приводим из-за ограничения объема статьи.

Таблица 1

	IMPF1	**PF**	**IMPF2**
Образцовые суффиксальные пары	—	*исключить*	*исключать*
	—	*открыть*	*открывать*
	—	*переписать*	*переписывать*
	[[*писать***]]**	*записать* <на пленку>	*записывать*
Суффиксальные пары с потенциальным НСВ1	[*лезть*]	*залезть* <на дерево>	*Залезать*
	[*рвать*]	*вырвать*<зуб>	*вырывать*
Образцовые Тройки	*капать*	*закапать* <капли в нос>	*закапывать*
	мазать	*намазать*	*намазывать*
	есть	*съесть*	*съедать*
	читать	*прочитать*	*прочитывать*
	гибнуть	*погибнуть*	*погибать*
	делить	*разделить*	*разделять*
	хватать	*схватить*	*схватывать*
	комкать	*скомкать*	*скомкивать*
Префиксальные пары с потенциальным НСВ2	*ломать*	*сломать* <ветку>	[*сламывать*]
	рисовать	*нарисовать*	[?*нарисовывать*]
Образцовые префиксальные пары	*строить*	*построить*	[??*постраивать*]
	делать	*сделать*	[???*сделывать*]*

* Ср. примеры из Интернета: *Она одевает короткую юбку и ботфотры, а губы* ***нарисовывает*** *себе до ушей; Ты веришь в судьбу? Тогда живи по ней! / Если нет, как я, тогда решай скорей! / Ведь если не веришь ни в Бога, ни в судьбу. /Тебе придется свою жизнь* ***постраивать*** *самому. На мой взгляд в начале самое главное научиться правильно* ***постраивать*** *алгоритмы. Для этого и Basic подойдет. Я вот сайт уже по CX4 начал* ***сделывать****. Неизвестно как, но к сессии курсовые и контрольные сами собой* ***сделываются****.*

Авторы выражают благодарность анонимным рецензентам, высказавшим ряд ценных замечаний, которые мы постарались учесть в окончательном варианте статьи.

Литература

1. *Апресян Ю. Д.* Трактовка избыточных аспектуальных парадигм в толковом словаре // Ю. Д. Апресян. Избранные труды. Т. II. М., 1995.
2. *Гиро-Вебер М., Микаэлян И.* Проблема видовых соотношений и ее отражение в аспектологическом словаре. // Ф. Леман (ред.) Глагольный вид и лексикография. Семантика и структура славянского вида IV. München: Otto Sagner verlag, 2006. P. 125–134.
3. *Гловинская М. Я.* Потенциальный видовой коррелят: жизнь и судьба // Aspekte, Kategorien und Kontakte slavischer Sprachen. Festschrift für Volkmar Lehmann zum 65. Geburtstag. Hamburg, 2008. S. 186–198.
4. *Зализняк Анна А., Микаэлян И.* Видовые тройки в русской аспектуальной системе // Русский язык: исторические судьбы и современность. III Международный конгресс исследователей русского языка. Москва, МГУ, 20–23 марта 2007. Труды и материалы. С. 463–464.
5. *Зализняк Анна А., Шмелев А. Д.* Введение в русскую аспектологию. М.: «Языки русской культуры», 2000.
6. *Зализняк Анна А., Шмелев А. Д.* Типы видовой связи // Труды международного семинара Диалог-2001 по компьютерной лингвистике и ее приложениям. Т. 1. Аксаково 2001.
7. *Зализняк Анна А., Шмелев А. Д.* О месте видовой пары в аспектуальной системе русского языка // Festschrift fur Peter Rehder zum 65. Geburtstag. Germano-savische Beiträge. B.21. Otto Sagner Verlag, Munchen, 2004.
8. *Зализняк Анна А., Микаэлян И. Л., Шмелев А. Д.* Видовая коррелятивность в русском языке: в защиту видовой пары // Вопросы языкознания, 2010, №1.
9. *Маслов Ю. С.* Вид и лексическое значение глагола в современном русском языке // Маслов Ю. С. Очерки по аспектологии. Л., 1984.
10. *Петрухина 2000* — Е. В. Петрухина. Аспектуальные категории глагола в русском языке в сопоставлении с чешским, словацким, польским и болгарским. М., 2000.
11. *Храковский 2005* — В. С. Храковский. Аспектуальные тройки и видовые пары // Русский язык в научном освещении. № 9, 2005.
12. *Цейтлин 2009* — Очерки по словообразованию и формообразованию в детской речи. М., 2009.
13. *Ясаи 2001* — Л. Ясаи. О специфике вторичных имперфективов видовых корреляций // Исследования по языкознанию: К 70-летию чл.-корр. РАН А. В. Бондарко. СПб., 2001.
14. *Guiraud-Weber M.* Le verbe russe. Temps et aspect. Aix-en-Provence, 2004.
15. *Janda L.* Aspectual clusters of Russian verbs // Studies in Language. V. 31, № 3, 2007. P. 607–648.
16. *Mikaelian I., Shmelev A., Zalizniak Anna A.* Imperfectivization in Russian // Meaning-Text Theory 2007, Proceedings of the 3rd International Conference on Meaning-Text Theory, Klagenfurt, May 20–24, 2007.

Анализ эффективности статистических методов выявления коллокаций в текстах на русском языке[1]

Study of effectiveness of statistical measures for collocation extraction on russian texts

Захаров В. П. (vz1311@yandex.ru),
Хохлова М. В. (vertikal-maria@yandex.ru)

Санкт-Петербургский государственный университет
Институт лингвистических исследований РАН, Санкт-Петербург

Аннотация. Описаны результаты исследования по выявлению устойчивых сочетаний в русском языке. Эксперимент заключался в нахождении и анализе биграмм с частотными глаголами и существительными русского языка. Цель — изучить сочетаемостные характеристики данных лексических единиц, соотнести результаты, полученные на основе различных мер ассоциации на разных корпусах, сравнить наиболее популярные меры ассоциации. Рассматриваются требования к программному обеспечению.

1. Понятие коллокации в лингвистике

Термин «коллокация», хотя и вошел в постоянное употребление сравнительно недавно, по праву занимает одно из ключевых мест в современной лингвистике. В широком смысле это комбинация двух или более слов, имеющих тенденцию к совместной встречаемости. Коллокации в настоящий момент играют ведущую роль в лексикографической практике (Atkins 2008; Hausmann 1979; Hausmann 1985; Kilgarriff 2006; Sinclair 1991). В последнее время за рубежом и в России создаются специальные словари коллокаций (Benson 1986; Crowther et al. 2002; Kjellmer 1994; Krishnamurthy et al. 2006; Sinclair 1995; Бирюк 2008; Денисов 2002, Кустова и др. 2008).

Однако существующие словари устойчивых словосочетаний, во-первых, охватывают далеко не полный их перечень, во-вторых, часто делают это недостаточно последовательно. Особенно это справедливо для русского языка. Поэтому актуальность работ по автоматическому выявлению коллокаций из текстов несомненна.

В настоящее время мы видим несколько важнейших прикладных задач, где есть нужда в автоматизированных методах извлечения коллокаций из больших корпусов текстов. В частности, это составление словарей и других лексикографических пособий, составление онтологий, обучение языку, отладка лингвопроцессоров, задачи информационного поиска.

Кратко коснемся самого понятия коллокация. Существуют различные определения этого понятия. В целом в основе большинства определений коллокации лежит явление семантико-грамматической взаимообусловленности элементов словосочетания (см., напр., (Иорданская, Мельчук 2007)).

Термин «коллокация» в русскоязычной научной литературе впервые появился в Словаре лингвистических терминов О. С. Ахмановой (Ахманова 1966). Первой работой в российской лингвистике, полностью посвященной исследованию понятия коллокации на материале русского языка, является монография Е. Г. Борисовой (Борисова 1995а).

В настоящее время термин «коллокация» нашел широкое применение в корпусной лингвистике, в рамках которой понятие коллокации переосмысливается или упрощается по сравнению с традиционной лингвистикой. Этот подход смело можно назвать статистическим. Во главу угла ставится частота совместной встречаемости, поэтому

[1] Данная работа выполнена при частичной поддержке гранта РФФИ 10-07-00563-а «Создание интегрированной автоматизированной системы для лингвистических исследований».

коллокации в корпусной лингвистике могут быть определены как ст*атистически устойчивые словосочетания*. При этом статистически устойчивое сочетание может быть как фразеологизированным, так и свободным. За последние годы появилось большое число исследований и разработок, посвященных коллокациям, затрагивающих как теоретические аспекты статистического подхода к данному понятию, так и практические методы выявления коллокаций.

И именно появление больших репрезентативных корпусов текстов позволяет получить достоверные данные о частоте того или другого сочетания в языке в целом. Высокая величина частоты совместной встречаемости, казалось бы, говорит об устойчивости данного сочетания. Однако этой характеристики недостаточно, чтобы говорить о предпочтительной сочетаемости тех или других слов. Поэтому был выработан целый ряд статистических мер (они получили название «меры ассоциации», или меры ассоциативной связанности, англ. association measures), вычисляющих силу связи между элементами в составе коллокации. В общем случае, эти меры учитывают как частоту совместной встречаемости, так и другие параметры, прежде всего частоту в данном корпусе каждого отдельного элемента.

Тем не менее, одних статистических данных недостаточно. Необходимо ответить на вопрос, каким еще требованиям должны соответствовать такие статистически устойчивые словосочетания.

2. Коллокации под углом зрения статистики

Практически большинство корпусных менеджеров обладают способностью производить подсчеты частот слов или словоформ и частот совместной встречаемости. Существует большое число мер ассоциации, которые основываются на этих данных. Общее количество этих мер исчисляется многими десятками. Значения мер ассоциации можно считать показателями силы синтагматической связи между элементами словосочетаний. Описание наиболее распространенных мер см. (Evert 2004). Чаще других используются MI, t-score и log-likelihood. Некоторые корпусные менеджеры предоставляют возможность вычисления этих мер.

Мера ***MI (mutual information)***, введенная в работе (Church, Hanks 1990), сравнивает зависимые контекстно-связанные частоты с независимыми, как если бы слова появлялись в тексте совершенно случайно:

$$MI(n,c)=log_2 \frac{f(n,c)\times N}{f(n)\times f(c)}, \text{где}$$

n — ключевое слово (node); c — коллокат (collocate); f(n,c) — частота встречаемости ключевого слова n в паре с коллокатом c; f(n), f(c) — абсолютные (независимые) частоты ключевого слова n и слова c в корпусе (тексте); N — общее число словоупотреблений в корпусе (тексте).

Если значение MI (n,c) больше определенного значения (для русского языка часто называется значение 3 и больше), тогда данное сочетание слов можно считать статистически значимым. Для MI (n,c) меньше нуля говорится, что *n* и *c* находятся в отношении дополнительной дистрибуции.

Есть различные модификации этой формулы, различными способами повышающие значение f(n,c) (Oakes 1998; Kilgarriff 2001).

Мера ***t-score*** также учитывает частоту совместной встречаемости ключевого слова и его коллоката, отвечая на вопрос, насколько не случайной является сила ассоциации (связанности) между коллокатами:

$$t-score=\frac{f(n,c)-\frac{f(n)\times f(c)}{N}}{\sqrt{f(n,c)}}$$

Также достаточно часто применяется мера, известная под названием ***log-likelihood,*** или *логарифмическая функция правдоподобия* (Dunning 1993).

$$log\text{-}likelihood = 2\sum_{ij} Oij \times \log\frac{Oij}{Eij}, \text{где}$$

Oij, Eij — наблюдаемая и ожидаемая частоты (подробнее см. (Evert 2004: 83)).

3. Цель и методы

Цель работы — сравнительный анализ различных мер ассоциации на основе корпусов русского языка. Кроме того, исследуется зависимость результатов (списка коллокаций, полученного на основе одной и той же меры) от текстового материала (тип текста). Работа выполнялась, в основном, на материале корпусов университета г. Лидс, составленных С. А. Шаровым на базе разных подмножеств Национального корпуса русского языка (НКРЯ), корпуса Интернет-текстов и др.

Сервис на сайте университета г. Лидс[2]позволяет выбрать одну из трех мер ассоциации (MI, t-score, log-likelihood) (варианты формул для вычисления выбраны С. А. Шаровым) или их совокупность, указать часть речи коллоката и расстояние между словами. Также имеется возможность проводить поиск коллокатов по лемме или словоформе. Не-

[2] http://corpus.leeds.ac.uk/ruscorpora.html

обходимо отметить, что каждый элемент в корпусе, включая знаки препинания, считается словом. Как следствие, среди результатов оказываются бессмысленные по сути комбинации, например, глаголов со знаками препинания.

Основным материалом исследования послужили 10 частотных глаголов русского языка: *быть*, *сказать*, *мочь*, *говорить*, *знать*, *стать*, *есть*, *хотеть*, *видеть*, *идти*. Эксперимент заключался в нахождении биграмм, одним из компонентов которых является глагол из приведенного списка.

4. Результаты

Результаты поиска коллокаций по корпусу НКРЯ-[3]были сведены в таблицы, представляющие собой объединение коллокаций, полученных на основе трех вышеуказанных мер (см. Табл. 1). Далее были удалены бессмысленные коллокации, т. е. комбинации глаголов со служебными словами и знаками пунктуации. Каждой коллокации был приписан свой ранг.

[3] Подмножество в 50 млн. словоупотреблений

Табл. 1. Часть таблицы результатов для глагола «***говорить***» (левый контекст) (модель Adv+V), отсортированных по мере MI

	Collocation	**Joint**	**Freq1**	**Rank MI**	**MI score (7,08–2,14)**	**Rank LL**	**LL score (1064,06–2,96)**	**Rank T-score**	**T- score (22,79–1,96)**
33.	честно говорить	527	2339	1.	7,08	1.	1064,06	101.	1,96
34.	постоянно говорить	62	4158	2.	7,04	14.	40,59	85.	2,26
35.	условно говорить	90	585	3.	6,53	8.	162,73	91.	2,11
36.	обиженно говорить	5	208	4.	6,46	77.	4,37	16.	6,52
37.	грубо говорить	130	988	5.	6,30	6.	224,23	93.	2,10
38.	умело говорить	23	2034	6.	6,20	33.	12,26	70.	2,40
39.	откровенно говорить	139	1203	7.	6,12	5.	230,24	94.	2,09
40.	собственно говорить	333	3114	8.	6,00	2.	538,32	99.	1,97
41.	жалобно говорить	6	481	9.	5,91	94.	3,45	25.	4,96
42.	……								

В столбцах таблицы, кроме самой коллокации, указываются следующие характеристики: *Joint — частота совместной встречаемости, Freq1 — частота коллоката (левый контекст), Rank MI — ранг по мере MI, MI score — значение меры MI, Rank LL — ранг по LL (log-likelihood), LL score — значение LL, Rank T-score — ранг по t-score, T-score — значение t-score*.

Анализ данных Табл. 1 (всего 101 коллокация) показывает, что ранги коллокаций, полученных на основе разных мер, не совпадают. Наиболее отличается мера t-score; меры MI и LL, наоборот, часто демонстрируют близкие ранги для найденных коллокаций (см. Табл. 1, строки 1, 3, 5, 7, 8), что верно и для других глаголов и синтаксических конструкций.

В словарных статьях толковых словарей для глагола «*говорить*» особенно отмечены устойчивые словосочетания, компонентом которых является форма деепричастия «*говоря*». Мы провели поиск биграмм, компонентом которых является именно данная форма слова. Ниже приведена сравнительная таблица для этого случая (см. Табл. 2).

Табл. 2. Частотные данные и меры ассоциации для глагола «***говорить***»
(первое значение для леммы /***второе значение курсивом*** для формы деепричастия)

Collocation	**Joint**	**Freq1**	**MI score**	**LL score**	**T score 1,96**
искренне говоря	12/***5***	1562	2,94/***4,92***	4,49/***6,11***	2,74/***2,16***
точно говоря	66/***41***	9893	2,64/***5,29***	21,09/***55,31***	2,21/***6,24***
просто говоря	214/***144***	28089	2,19/***5,60***	79,38/***209,98***	2,02/***11,75***
откровенно говоря	139/***104***	1203	6,12/***9,67***	230,24/***299,54***	2,09/***10,19***
честно говоря	527/***429***	2339	7,08/***10,98***	1064,06/***1690,55***	1,96/***22,33***
объективно говоря	7/***6***	502	4,24/***6,82***	4,37/***11,22***	4,16/***2,43***
образный говоря	51/***44***	233	3,00/***10,80***	102,07/***145,01***	2,32/***6,63***
строгий говоря	166/***146***	4239	4,55/***8,34***	184,16/***351,80***	2,08/***12,05***
условно говоря	90/***87***	585	6,53/***10,45***	162,73/***275,55***	2,11/***9,32***
грубо говорить	130/***127***	988	6,30/***10,24***	224,23/***392,55***	2,10/***11,26***
мягко говоря	252/***247***	3916	5,27/***9,22***	341,77/***672,86***	2,01/***15,69***
коротко говоря	267/***265***	6540	4,61/***8,58***	301,73/***662,08***	1,97/***16,24***
собственно говоря	333/***332***	3114	6,00/***9,97***	538,32/***996,77***	1,97/***18,20***
упрощенно говоря	5/***5***	34	3,07/***10,44***	8,92/***15,78***	7,53/***2,23***

Табл. 3. Значения мер ассоциации для коллокаций со словом ***война,*** совпавших с коллокациями словаря Е. Г. Борисовой

Collocation	Joint	Freq1	Freq2	LL score	MI	T-score
вспыхивать война	5	1201		**8,29**	**6,20**	**2,21**
идти война	167	47464		**264,43**	**5,96**	**12,72**
кровопролитный война	6	251		**15,18**	**8,72**	**2,44**
разражаться война	9	881		**18,94**	**7,50**	**2,98**

Табл. 4. Меры ассоциации для коллокаций со словом война по БАС-17

Collocation	Joint	Freq1	Freq2	LL score	MI	T-score
гражданский война	194	12469		**451,11**	**8,10**	**13,88**
локальный война	8	860		**16,46**	**7,36**	**2,81**
мировой война	154	25171		**285,92**	**6,76**	**12,29**
партизанский война	45	728		**135,77**	**10,09**	**6,70**
холодный война	171	4747		**469,90**	**9,31**	**13,06**

Анализ результатов показывает, что в этом случае коллокации обладают значительно бóльшим числовым значением для всех мер ассоциации. См., например, строки «точно говоря», «откровенно говоря». Из этого можно сделать вывод, что иногда статистические меры для поиска коллокаций следует применять к словоформам, а не к леммам.

Далее мы исследовали зависимость состава и ранжирования списков коллокаций, полученных на основе одной и той же меры MI на разных корпусах текстов русского языка (НКРЯ (117 млн. словоупотреблений) и газетный корпус (70 млн.)), то есть зависимость списка коллокаций от типа текста.

Анализ коллокаций, полученных на этих двух корпусах, показывает, что грубо их можно разбить на две части: присутствующие в обоих корпусах (часто с близкими рангами) и присутствующие только в одном из них. Видимо, это говорит о принадлежности коллокатов, в данном случае, наречий, выданных только по одному из корпусов, к определенному жанру. И действительно, анализ контекстов употребления наречий *пространно, модно, полусерьзно, фигурально* показывает преобладание их в корпусе художественных текстов. Но еще более разительную картину дает сравнений коллокаций, полученных на основе НКРЯ (117 млн. словоупотреблений) и Интернет-корпус (188 млн.), где из 13 первых коллокаций из НКРЯ, отсортированных по мере MI, в Интернет-корпусе присутствует только одна. Это говорит о том, что для разных жанров, возможно, следует применять разные меры.

Также следует учитывать, что разные меры поразному реагируют на частоту слов, образующих коллокацию, и на частоту совместной встречаемости. Так утверждается, что MI чувствительна к низкочастотным словам, а t-score полезна для нахождения высочастотных коллокаций (Evert 2004; Čermák 2006; Kilgarriff 2006).

Мы провели сравнение коллокаций, полученных автоматически на основе разных мер ассоциации, с данными различных словарей (более подробно см. (Хохлова 2008)). Материалом послужили коллокации 19 существительных, не имеющих омонимов (по Малому академическому словарю (МАС) (Словарь русского языка 1981–1984)) и представленных в словаре коллокаций русского языка Е. Г. Борисовой (Борисова 1995b). Исследование проводилось также на базе корпусов русских текстов, созданных в университете г. Лидс. Были проанализированы коллокации на базе газетного корпуса (78 млн. слов).

Результаты запроса для каждого существительного (выявленные коллокации) сравнивались со словарными статьями, приведенными для этих существительных в Словаре коллокаций (Борисова 1995b), в толковых словарях русского языка: БАС-17 (Словарь современного русского языка 1948–1965) и МАС (Словарь русского языка 1981–1984) — и в Словаре синонимов и сходных по смыслу выражений (Абрамов 2006). Приведем некоторые результаты для слова *война* (по первым 100 биграммам для левого контекста) (Таблица 3).

Будем называть коллокации, приведенные в словаре Е. Г. Борисовой и входящие в таблицы, «правильными».

В таблице 4 приведены словосочетания, найденные на слово *война* в БАС-17.

На рис. 1 приводится график для БАС-17 (значения меры MI по оси ординат и ранги биграмм по оси абсцисс). Темным цветом обозначены «правильные» коллокации из словаря Борисовой (ранги 12, 18, 38, 41) и дополнительные коллокации, найденные в БАС-17 (ранги 2, 8, 13, 34, 44).

Мы видим, что «правильные» коллокации Борисовой и устойчивые словосочетания БАС распределены в левой половине шкалы.

На таком же графике для БАС-17 по значениям меры log-likelihood «правильные» коллокации из словаря Борисовой (ранги 5, 35, 43 и 60) и дополнительные коллокации, найденные в БАС-17 (ранги

1, 2, 3, 5, 8), также находятся в левой части графика. Для словосочетаний, найденных по мере t-score, мы также наблюдаем «скученность» «словарных» коллокаций в левой половине графика.

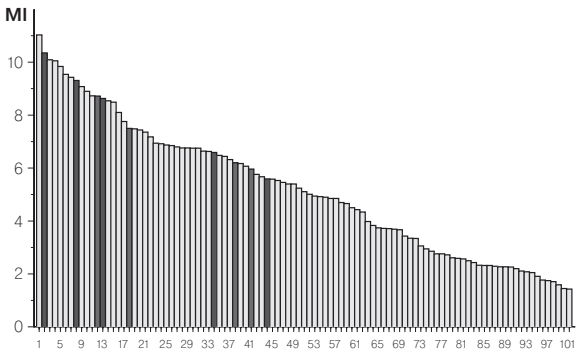


Рис. 1. Значения меры MI для коллокаций со словом *война*

Для всех полученных сочетаний наблюдается одинаковая тенденция: чем меньше значение меры, тем больше вероятность, что эти словосочетания не зафиксированы как устойчивые в словарях русского языка. Таким образом, можно сказать, что данные о сочетаемости, приведенные в словарях, совпадают с данными, полученными на основе мер ассоциации.

Важным представляется и тот факт, что в результате эксперимента были выделены сочетания, не зафиксированные ни в одном из словарей. Анализ подобных сочетаний показал, что биграммы, находящие на самом верху списка (отсортированного по убыванию по одной из мер), с некоторой долей вероятности оказываются устойчивыми и, следовательно, могут быть внесены в словарь.

Как уже было сказано, поверх статистических критериев должны работать и другие методы, основывающиеся на собственно лингвистических моделях. Данная идея заложена и реализована в известной системе Sketch Engine (Kilgarriff et al. 2004). Она выдает для заданного ключевого слова типичные словосочетания, обусловленные, с одной стороны, синтаксисом, накладывающим ограничение на сочетаемость слов в заданном языке, а с другой стороны, вероятностными закономерностями, связанными с семантикой и языковым узусом. Результат работы программы представлен наиболее устойчивыми словосочетаниями с учетом грамматических (структурных) формул. Однако и с помощью обычных корпусных менеджеров также можно получить похожие результаты (см. Табл. 5, поиск на сайте университета г. Лидс).

Результаты поиска и выдачи коллокаций в таком виде удобны для лексикографов, которые могут выбрать для словарей примеры разного типа, и для лингвистов, изучающих лексику и синтаксис в определенном аспекте.

Табл. 5. Результаты поиска коллокаций с глаголом «*сказать*» (отсортированные по мере LL)

V + Adv (Pred)		V + N	
Collocation	**LL**	**Collocation**	**LL**
сказать трудно	56,37	сказать правда	31,68
сказать нельзя	20,33	сказать слово	14,13
сказать точно	18,34	сказать гадость	6,44
сказать вслух	17,65	сказать комплимент	4,73
сказать особо	16,90	сказать неправда	4,60
сказать честно	12,90	сказать тост	3,92
можно сказать	3274,07	хотеть сказать	2550,77
надо сказать	1768,42	хотеться сказать	247,53
нельзя сказать	524,47	успевать сказать	81,72
трудно сказать	518,11	следовать сказать	70,24
точно сказать	183,63	забывать сказать	68,04
тужно сказать	145,33	смочь сказать	47,21

5. Выводы

При сравнении сочетаний, полученных с помощью статистических методов, со словарями наблюдается одинаковая тенденция: чем меньше значение меры, тем больше вероятность, что эти словосочетания не зафиксированы как устойчивые в словарях русского языка, и наоборот. Большинство коллокаций, зафиксированных в словарях, оказывается в верхней части списка, составленного на основе одной из мер ассоциации. Таким образом, можно сказать, что данные об устойчивой сочетаемости, приведенные в словарях, совпадают с данными, полученными на основе мер ассоциации, или, по-другому, что статистические меры ассоциации достаточно хорошо выявляют реально существующие семантико-синтагматические связи.

Сравнительный анализ различных мер ассоциации, проведенный по совокупности всех данных, полученных нами для разных частей речи, показывает следующее.

Мера MI, возможно, дает наилучшие усредненные результаты. Она позволяет выделить устойчивые фразеологизированные словосочетания, а также сочетаниями, где в качестве коллокатов выступают имена собственные, а также низкочастотные специальные термины. К недостаткам использования меры t-score можно отнести то, что она, в первую очередь, выделяет коллокации с очень частотными словами-коллокатами, в частности, со служебными словами. Поэтому для t-score необходимо задавать список стоп-слов, чтобы «отбросить» самые частотные слова, сочетания с которыми неизменно оказываются вверху таблицы: предлоги, местоимения или союзы. Впрочем, это, видимо, справедливо и для других мер. Это следует из наших экспериментов, это же подтверждается и в других публикациях (см., напр., Baroni 2008; Evert 2004; Čermák 2006;

Khokhlova 2009b; Kilgarriff 2006; Křen 2006; Пивоварова 2010; Хохлова 2008).

Возможно, что стоит изучить возможности объединения разных мер, например, ввести величину, равную сумме их рангов (Cvrček 2006).

Остается открытым вопрос, стоит ли учитывать в статистических мерах при поиске коллокаций леммы или словоформы (см. Табл. 2).

Следует также принимать во внимание структурные синтаксические формулы и семантические ограничения, которые лежат в основе коллокаций. Их комбинация со статистическими подходами, по нашему мнению, может дать неплохие результаты (Khokhlova 2009a).

В качестве общего вывода из проведенного исследования мы хотим отметить, что существующий программный инструментарий автоматического выявления коллокаций на основе статистических методов весьма неудовлетворителен — как в части лингвистического обеспечения, так и с точки зрения выходных интерфейсов, и его следует развивать. В первую очередь, важно уметь находить разрывные коллокации со свободным порядком, искать коллокаты не только по леммам, но и по словоформам, искать коллокаты для гнезда опорных однокоренных слов, уметь варьировать размер окна, в котором ищутся коллокаты. Нередко реальные коллокации представляют собой n-граммы, где n больше двух, тогда встает вопрос выбора формул для мер ассоциации для таких словосочетаний. При выборке коллокаций из текста должна производиться обработка знаков препинания и служебных слов, имен собственных и т. п.

Особое значение имеет выдача коллигаций — коллокаций, построенных по определенной синтаксической модели, учет отношения зависимости между элементами коллокаций.

Литература

1. *Абрамов Н. М.* Словарь русских синонимов и сходных по смыслу выражений. М.: 2006.
2. *Ахманова О. С.* Словарь лингвистических терминов, М.: 1966.
3. *Бирюк О. Л., Гусев В. Ю., Калинина Е. Ю.* Словарь глагольной сочетаемости непредметных имен русского языка. М., 2002. (См. // http://dict.ruslang.ru/)
4. *Борисова Е. Г.* Коллокации. Что это такое и как их изучать. М.: 1995. (Борисова 1995a).
5. *Борисова Е. Г.* Слово в тексте. Словарь коллокаций (устойчивых словосочетаний) русского языка с англо-русским словарем ключевых слов. М.: 1995. (Борисова 1995b).
6. *Денисов П. Н., Морковкин В. В.* Словарь сочетаемости слов русского языка. М., 2002.
7. *Иорданская Л. Н., Мельчук И. А.* Смысл и сочетаемость в словаре. М.: 2007.
8. *Кустова Г. И.* Словарь русской идиоматики. Сочетания слов со значением высокой степени. // http://dict.ruslang.ru/
9. *Пивоварова Л. М.* Подводные камни статистических мер (в печати). 2010.
10. *Словарь* русского языка: В 4 т.: 1981–1984, Т. 1–4, Москва. (МАС)
11. *Словарь* современного русского литературного языка: В 17 т.: 1948–1965, Т. 1–17, Москва, Ленинград. (БАС-17).
12. *Хохлова М. В.* Экспериментальная проверка методов выделения коллокаций // Slavica Helsingiensia 34. Инструментарий русистики: Корпусные подходы. Под ред. А. Мустайоки, М. В. Копотева, Л. А. Бирюлина, Е. Ю. Протасовой. Хельсинки: 2008. С. 343–357.
13. *Atkins S. and Rundell M.* The Oxford Guide to Practical Lexicography. Oxford University Press, 2008.
14. *Baroni M., Evert S.* Statistical methods for corpus exploitation. // In A. Lüdeling and M. Kytö (eds.), Corpus Linguistics. An International Handbook, article 36. Mouton de Gruyter, Berlin: 2008.
15. *Benson M.* (Ed.) The BBI combinatory dictionary of English. Amsterdam: John Benjamins Publishing Co, 1986.
16. *Čermák F.* Metoda zjistovani kolokacni platnosti frekventovanych bigramu pomoci ranku. // Čermák F. (ed.) Kolokace. Ústav Českého národního korpusu, Praha: 2006. P. 94–105.
17. *Church K., Hanks P.* Word association norms, mutual information, and lexicography // Computational Linguistics, 1996, № 16(1), P. 22–29.
18. *Crowther J., Dignen S. & Lea D.* (Eds.). Oxford Collocations Dictionary for Students of English. Oxford: Oxford University Press, 2002.
19. *Cvrček V.* Metoda zjišťování kolokační platnosti frekventovaných bigramů pomocí ranku. // Čermák F. (ed.) Kolokace. Ústav Českého národního korpusu, Praha: 2006. P. 36–55.
20. *Dunning T.* Accurate Methods for the Statistics of Surprise and Coincidence. // Computational Linguistics. 1993. Volume 19, №1, P. 61–74.
21. *Evert S.* The Statistics of Word Cooccurences Word Pairs and Collocations. PhD thesis. Institut für Maschinelle Sprachverarbeitung (IMS), Universität Stuttgart: 2004.
22. *Khokhlova M.* Applying Word Sketches to Russian. // In Proceedings of Raslan 2009. Recent Advances in Slavonic Natural Language Processing. Brno: Masaryk University, P. 91–99. (Khokhlova 2009a)
23. *Khokhlova M., Zakharov V.* Statistical collocability of Russian verbs // After Half a Century of Slavonic Natural Language Processing. Dana Hlaváčková, Aleš Horák, Klára Osolsobě, Pavel Rychlý (Eds.). Brno, 2009. P. 105–112. (Khokhlova 2009b)
24. *Kilgarriff A.* Collocationality (and how to measure it) // Proceedings of the Euralex International Congress. Torino, 2006.
25. *Kilgarriff A.* Web as Corpus. // Proceedings of Corpus Linguistics. Lancaster, UK. 2001.
26. *Kilgarriff A., Rychly P., Smrz P., Tugwell D.* The Sketch Engine // Proceedings of the Eleventh EURALEX International Congress. Lorient, 2004. P. 105–116.
27. *Kjellmer G.* A dictionary of English collocations: based on the Brown corpus : in three volumes. Oxford; New York: Clarendon Press: Oxford University Press, 1994.
28. *Křen M.* Kolokacni miry a cestina: srovnani na datech ČNK. // Čermák F. (ed.) Kolokace. Ústav Českého národního korpusu, Praha: 2006. P. 223–248.
29. *Krishnamurthy R. & Keith B.* 2006. Collocations Encyclopedia of Language & Linguistics. Oxford: Elsevier. P. 596–600.
30. *Oakes M.* Statistics for Corpus Linguistics. Edingurgh: 1998.
31. *Sinclair J.* Collins COBUILD English collocations on CD-ROM: Harper Collins, 1995.
32. *Sinclair J.* Corpus, concordance, collocation. Oxford: Oxford University Press, 1991.

Местоимения с кванторным антецедентом в русском языке

Pronouns with quantified antecedents in russian

Ивлиева Н. В. (natasha.ivlieva@gmail.com)
MIT

Настоящая работа посвящена исследованию ограничений на анафорическую связь между кванторным словом или квантифицированной именной группой и местоимением 3 лица *он* (а также посессивным местоимением *его*) в русском языке.

1. Введение

Целью настоящей работы является установление ограничений на употребление местоимений с кванторным антецедентом, действующих в русском языке.

С точки зрения семантики анафора с кванторным антецедентом представляет особый интерес, поскольку кванторные выражения не имеют референции, а значит, задействованные принципы интерпретации отличаются от тривиального отождествления местоимения и его антецедента с одним референтом. Сравним предложения (1) и (2):

(43) Иван думает, что он гений.

(44) Каждый человек думает, что он гений.

Мы можем сказать, что в (1) местоимение представляет собой референтное выражение, которое обозначает того же самого индивида, что и его антецедент. Иными словами, предложение (1) может быть заменено следующей синонимичной перифразой:

(45) Иван думает, что Иван гений.

Анафорическая связь в таком случае называется *кореферентностью*. Антецедент и местоимение кореферентны — обозначают один и тот же предмет.

С предложением (2) ситуация несколько другая — местоимение *он* в этом случае обозначает не какой-то фиксированный предмет, однако, интерпретация местоимения находится в зависимости от кванторного выражения. Значение, которое принимает местоимение, варьирует в зависимости от выбора конкретного студента. Таким образом, можно сказать, что в данном случае местоимение является естественноязыковым коррелятом связанной переменной логики первого порядка. Интерпретация предложения (2) может быть представлена следующим образом:

(46) для каждого студента x (x думает, что x гений).

Такой тип связи между кванторным выражением и местоимением получил название *связывание переменной*. Местоимение в данном случае выражает переменную, которую связывает квантор.

С точки зрения синтаксиса анафора с кванторным антецедентом — также исключительно интересное явление, поскольку на нее накладываются более жесткие ограничения, чем на анафору с неквантифицированными антецедентами.

Чтобы показать это, приведем пару примеров из русского языка:

(47) *Когда ему$_i$ подарили подарок, Вася$_i$ обрадовался.*

(48) **Когда ему$_i$ подарили подарок, каждый$_i$ обрадовался.*

Как видно из примеров (5) и (6), местоимение *ему* может анафорически отсылать к референтной ИГ *Вася*, но не может отсылать к кванторному выражению *каждый*.

На материале русского языка попытки установить ограничения на анафору с кванторным ан-

тецедентом предпринимались разве что в работах Е. В. Падучевой [Падучева 1985] и К. И. Казенина [Казенин 2000]. Стоит заметить, однако, что в этих работах ограничения установлены на явно недостаточном языковом материале, и, как будет показано ниже, они не являются вполне верными.

Рамки исследования были намеренно сужены. Рассматриваются лишь случаи связывания местоимений универсальным кванторным словом *каждый* и именными группами, в которых это слово является модификатором вершины. Уже этих данных достаточно для того, чтобы обозначить круг ограничений, релевантных для проблемной области в целом.

Также отметим, что в настоящей работе нас будут интересовать только случаи связывания квантором местоимения в форме единственного числе. Такое сужение круга описываемых явлений не случайно: на кореферентность квантора местоимению во множественном числе действуют более слабые ограничения, ср. (7).

(49) *Все тяготы, с которыми сталкивался [каждый член группы]$_i$ […] воспринимались ими$_i$/*им$_i$ как испытание силы их веры.* [В. Н. Павленко, К. Ваннер. Особенности психологии евангельских христиан-баптистов // «Вопросы психологии», №5, 2004]

Как было замечено в работе [Reinhart 1983a] квантифицированная ИГ, находясь в одном предложении, может контролировать референцию местоимения во множественном числе в другом предложении, иными словами, контролировать не синтаксическую, а дискурсивную анафору, что невозможно в случае с местоимениями в единственом числе. Ср. русские примеры, аналогичные английским из [Reinhart 1983a]:

(50) *Я поговорил с [каждым студентом]$_i$ о предстоящей защите. Они$_i$ очень волнуются.*

(51) **Я поговорил с [каждым студентом]$_i$ о предстоящей защите. Он$_i$ очень волнуется.*

В предложениях (7–8) употребление местоимения во множественном числе обусловлено тем, что референция происходит ко всему множеству (к которому «отсылает» квантор) в целом. При такой референции ограничения на кванторно-местоименную анафору существенно отличаются от тех, которые накладываются при *семантическом связывании* квантором местоимения (когда местоимение семантически выражает переменную, связанную квантором). В данной же работе, как уже было сказано, нас будут интересовать только специфические ограничения на *связывание* квантором местоимения.

2. Линейные и структурные ограничения

Обычно ограничения на употребление местоимений с кванторными антецедентами формулируются либо в терминах линейного предшествования: квантор должен линейно предшествовать местоимению, чтобы его связать (Postal 1972, Chomsky 1976 и др.), либо в терминах структурного приоритета (c-command): квантор должен иметь структурный приоритет над местоимением в структуре составляющих (Reinhart 1983 и др.)[1].

В этом разделе мы попытаемся проверить, действуют ли эти ограничения в русском языке.

То, что в русском языке квантор должен линейно предшествовать местоимению, которое он связывает, отмечалось в работах [Падучева 1985, Reuland&Avrutin 2005] и др. В НКРЯ не нашлось примеров, где бы связанное местоимение предшествовало квантору. Безусловно, это не может свидетельствовать о неграмматичности предложений, где условие линейного предшествования не выполняется. Тем не менее ниже приводится несколько показательных примеров из корпуса.

(52a) *Каждый$_i$, поскольку его$_i$ воля нравственна, внутренно участвует в этой всеобщей организации нравственности…* [В. Соловьев. Оправдание добра (1894–1899)]

(10b) **Поскольку его$_i$ воля нравственна, каждый$_i$ внутренно участвует в этой всеобщей организации нравственности.*

(53a) *Надо дать каждому$_i$ в руки его$_i$ деньги за ОМС.* [Елена Костюк. Здоровый популизм доведет народ… // «Время МН», 2003.05.26]

(11b) **Его$_i$ деньги за ОМС каждому$_i$ надо дать в руки.*

(54a) *[Каждую комету]$_i$ по ее$_i$ циклам можно проследить.* [Павел Глоба. Лекция, прочитанная в Минске (2001.10.26)]

(12b) **По ее$_i$ циклам каждую комету$_i$ можно проследить.*

В примерах (10a), (11a) и (12a) условие линейного предшествования квантора выполняется, а в (10b), (11b) и (12b) — нет. Возможно, именно этим объясняется их неграмматичность, однако, как

[1] В работах [Bresnan 1994, 1998] ограничения на связывание местоимений кватором представляют собой продукт взаимодействия линейных и структурных ограничений, ср. также принцип *precede and command* в [Langacker 1969].

было замечено в работе [Казенин 2000], подобный контраст можно объяснить и в терминах структурного приоритета. Структурный приоритет квантора над местоимением в (10a) очевиден. Также при некоторых допущениях (о структуре трехместных глаголов и конструкции OV) структурный приоритет квантора можно постулировать в (11a) и (12a). При тех же допущениях структурного приоритета квантора над местоимением нет в (10b), (11b) и (12b). Таким образом, приведенные данные одинаково успешно объясняются двумя конкурирующими теориями, однако, если взглянуть на другие примеры, становится понятно, что ни одна из них не делает верных предсказаний.

Существуют примеры, в которых местоимение предшествует квантору, причем подавляющее большинство опрошенных носителей русского языка признают возможность связанной интерпретации:

(55) *Женщину, с которой он$_i$ мог бы поговорить по душам, ищет [каждый мужчина]$_i$.*

(56) *О любви, которая перевернет всю ее$_i$ жизнь, мечтает [каждая женщина]$_i$.*

(57) *В отношении к тому, что его$_i$ окружает, выражается и воплощается личное достоинство каждого$_i$.*

(58) *О том, что его$_i$ ждет смерть, [каждый заговорщик]$_i$ знал наперед.*

Эти примеры объединяет то, что в каждом из них изменен базовый порядок слов (SVO). В примере (13) прямое дополнение, в составе которого употреблено местоимение *он*, вынесено в начало предложения. В примерах (14–16) актантные предложные группы также находятся в начале предложения. В терминах порождающей грамматики эти предложения являются результатом *передвижения* составляющих. В исходных структурах этих предложений условие линейного предшествования квантора местоимению выполняется.

Возможно, объяснить грамматичность предложений (13–16) можно, используя понятие *реконструкции* (reconstruction, см. например, [Huang 1993, Chomsky 1995]). Реконструкцией называется процесс, вследствие которого элемент, подвергшийся передвижению, на уровне семантической интерпретации "возвращается" в свою исходную позицию. Иными словами, семантический компонент может просто «не заметить» синтаксического передвижения. Если так, то с точки зрения установления отношений связывания, например, предложение (13) может ничем не отличаться от предложения (17):

(59) *[Каждый мужчина]$_i$ ищет женщину, с которой он$_i$ бы мог поговорить по душам.*

В (17) связывание установлено в соответствии с принципом линейного предшествования. Поскольку (17) формально отличается от (13) только порядком слов и, предположительно, точнее отражает общую «исходную» структуру этих предложений, связывание возможно и в (13). Таким образом, грамматичность предложений (13–16) будет являться следствием действия все того же линейного ограничения (см. обсуждение строгой формулировки этого ограничения с теоретической точки зрения в [Ivlieva 2009]).

В предложениях типа (6) реконструкции не происходит: там местоимение входит в состав адъюнкта, у которого нет фиксированного положения в исходной структуре, вследствие чего он не подвергается реконструкции.

Впрочем, условие линейного предшествования является необходимым, но не достаточным условием для установления отношения связывания между квантором и местоимением в русском языке. Это показывают примеры типа (18), в которых это условие соблюдено, однако связывание невозможно.

(60) **Человек, который является хозяином [каждой собаки]$_i$, делает ей$_i$ прививки.* (пример из [Падучева 1985])

Неграмматичность (18) можно было бы объяснить действием условия структурного приоритета. Напомним, что условие структурного приоритета было предложено в работе Т. Рейнхарт [Reinhart 1983]. Оно заключается в том, что квантор, для того чтобы связать местоимение, должен иметь над ним структурный приоритет, причем имеется в виду, что это требование должно выполняться в поверхностной структуре, а не на уровне семантической интерпретации, например (как предлагалось во многих работах, ср. подъем квантора в [May 1977]). Рассмотрим подробнее возможность применения этого требования к данным русского языка.

Если квантор имеет структурный приоритет над местоимением 3 лица и при этом не является для него локальным подлежащим, связывание всегда возможно, см. примеры ниже.

(61) *На правах бывшего комсорга Люся Огородникова заняла преподавательское место и потребовала, чтобы каждый$_i$ отчитался, как он$_i$ живет.* [Даниил Гранин. Искатели (1954)]

(62) *Мы уже катаемся не вместе, а встали в пару, где каждый$_i$ вытворяет, что он$_i$ хочет.* [Наталья Бестемьянова, Игорь Бобрин, Андрей Букин. Пара, в которой трое (2000–2001)]

(63) *Каждый$_i$ выбирает то, что ему$_i$ ближе.* [Навстречу новому тысячелетию // «Мурзилка», №1, 2000]

(64) *ПК «Строительно касса» гарантирует [каждому своему члену]$_i$ направление его$_i$ паевого взноса только на приобретение недвижимости.* [Потребительский кооператив // «Пермский строитель», 2003.05.12]

Однако квантор далеко не всегда имеет структурный приоритет над местоимением, которое он связывает. Так, в предложениях (23–24) квантор находится в составе приименного генитивного зависимого и поэтому очевидным образом не может иметь структурного приоритета над местоимением:

(65) *Суд каждого$_i$ внутри него$_i$, в его$_i$ собственной душе — и ты присутствуешь на этом суде, это твое право.* [Олег Глушкин. Письмо для Бога (1990–1999)]

(66) *Специфику [каждой конкретной ситуации]$_i$ составляют ее$_i$ участники, обладающие индивидуальными свойствами, которые в данной комбинации не повторяются ни в одной другой ситуации.* [В. С. Храковский. Понятие сирконстанта и его статус (1999)]

В предложениях (25–26) квантор находится в составе предложной группы:

(67) *Но [после [каждого писателя]$_i$] остаются его$_i$ книги.* [Георгий Иванов // «Трамвай», №9, 1990]

(68) *Не* пора *ли вернуться к опыту дореволюционной России, когда [за [каждую реформу]$_i$] полностью отвечал ее$_i$ автор: за финансовую — Витте, за земельную — Столыпин…* [Александр Авсеевич, Евгения Дылева. Бедность — черта или замкнутый круг // «Петербургский Час пик», 2003.09.03]

Структурного приоритета нет и в (27–28), где квантор находится в зависимой клаузе, а местоимение — в главной:

(69) *Если [каждый полицейский]$_i$ будет знать, что его будут таскать для допросов и расследований, то он$_i$ в следующий раз постарается ничего не видеть, не слышать и ничего не знать.* [Раддай Райхлин. Как захватить власть // «Лебедь» (Бостон), 2003.08.04]

(70) *Очевидно, если мы хотим снабдить [каждого пользователя]$_i$ своей СУБД, то необходимо обеспечить ему$_i$ доступ на равных правах с другими пользователями.* [А. Б. Барский. Применение SPMD-технологии при построении сетевых баз данных с циркулирующей информацией // «Информационные технологии», 2004.07.26]

Таким образом, условие структурного приоритета является достаточным, но не необходимым условием для установления отношения связывания между квантором и местоимением в русском языке.

Попытка определить возможные конфигурации, в которых могут находиться местоимение и его кванторный антецедент, и в линейных, и в структурных терминах была предпринята в [Падучева 1985]. Как было сказано выше, в этой работе постулируется линейное ограничение. Кроме того, там говорится, что для того, чтобы квантифицированная ИГ могла связать местоимение, они должны находиться в определенных отношениях в синтаксической структуре, правда, при этом имеется в виду не структура составляющих, а структура зависимостей:

«Требуемое синтаксическое соотношение между ИГ — антецедентом ($ИГ_а$) и ИГ — местоимением ($ИГ_м$) состоит в том, чтобы $ИГ_а$ подчинялась сказуемому предложения (непосредственно), а $ИГ_м$ подчинялась тому же сказуемому через любое число промежуточных слов, в том числе союзов и относительных слов, которые вводят новое подчиненное предложение. В этом случае $ИГ_а$ контролирует $ИГ_м$».

Например, отсутствие кореферентности в предложении (29) объясняется так: антецедентная ИГ не контролирует местоимение, потому что между ними не соблюдается требуемое синтаксическое соотношение, а именно местоимение непосредственно подчинено сказуемому (*доволен*), а антецедент входит в состав придаточного, зависящего от этого сказуемого.

(71) *Если каждый$_i$ одержал победу, он$^*_i$ доволен.*

Однако анализ Падучевой не может объяснить все факты, включая уже приведенные. Например, он не объясняет грамматичность предложений (23–28).

3. Кванторные антецеденты в зависимых клаузах

В литературе предлагалось и такое ограничение на связывание местоимений кванторами: местоимение в главной клаузе не может быть связано кванторным антецедентом из зависимой клаузы.

По-видимому, оно действует и в русском языке, ср. невозможность связывания в (29), а также (30–31).

(72) *Помимо завистников и недоброжелателей, которые есть у каждого$_i$, на него$^*_i$ ополчились и кое-кто из китов.* [Еремей Парнов. Третий глаз Шивы (1985)]

(73) *В эту секунду она поняла, что та любовь, о которой мечтает [каждая женщина]$_i$, прошла мимо нее$^*_i$.* [А. И. Куприн. Гранатовый браслет (1911)]

В (29) местоимение находится в главном предложении, а квантор — в подчиненном ему обстоятельственном. В (30), как и в (31), местоимение опять же находится в главном предложении, а квантор — в относительном придаточном. Подобная конфигурация не допускает установления связывания.

Однако далеко не всегда нахождение квантора внутри придаточного предложения ведет к недопустимости связанной интерпретации. Ср. грамматичные предложения (32–33), в которых квантор находится в обстоятельственном придаточном, а связанное местоимение — в главном предложении[2]:

(74) *А потом добавил, что хотя [каждый случай]$_i$ уникален, он$_i$ всегда напоминает остальные.* [Нодар Джин. Учитель (1980–1998)]

(75) *Если кажд**ый**$_i$ будет заниматься своим делом, он$_i$ принесет больше пользы.* [Ирина Подлесова. Большинство колебалось. В воскресенье москвичи выбирали городскую думу // «Известия», 2001.12.16]

Встает вопрос, в каких именно контекстах возможно связывание квантором, находящимся в зависимой клаузе, местоимения в главной клаузе.

На первый взгляд кажется, что связывание возможно в тех случаях, когда возможна квантификация ситуаций — когда предложение содержит оператор, квантифицирующий ситуации (генерический оператор).

Заметим, что в предложениях типа (34) интерпретация со связыванием недоступна, т. к. в них обстоятельственные придаточные вводят единственный темпорально-ситуационный референт. Видимо, это связано с употреблением в обстоятельственном придаточном глагола в форме прошедшего времени совершенного вида[3].

(76) *Если *[каждый бизнесмен]$_i$ вложил деньги в строительство школы, он$_i$ сделал доброе дело.*

Однако такой анали не в состоянии объяснить неграмматичность предложений (35–37):

(77) *Если *каждый преподаватель задерживается на работе, он не ужинает.*

(78) *Если *каждый$_i$ студент пьяный, он$_i$ улыбается.*

(79) *Если *[каждый бизнесмен]$_i$ вкладывал деньги в строительство школы, он$_i$ делал доброе дело.*

Предложения (35–37) представляют собой генерические высказывания (предложения, содержащие генерический оператор), однако, не допускают связанной интерпретации (ср. с генерическими предложениями без квантора *каждый*: *Если служащий задерживается на работе, он не ужинает*).

Проверим, возможно ли связывание в предложениях с будущим временем и сослагательным наклонением. Примеры (38–39) показывают, что в предложениях с будущим временем и с сослагательным наклонением связывание возможно:

(80) Если *каждый$_i$ будет заниматься своим делом, он$_i$ принесет больше пользы.*

(81) Если *бы каждый$_i$ занимался своим делом, он$_i$ бы приносил больше пользы.*

Однако все опять не так тривиально. Одного наличия будущего времени или сослагательного наклонения недостаточно для возможности связывания. Ср. грамматичное (38) с неграмматичным (40) и контраст между (a) и (b) в (41):

(82) *Если *каждый$_i$ будет заниматься своим делом, президент вручит ему$_i$ награду.*

(83a) Если *[каждый бизнесмен]$_i$ вложит деньги в строительство школы, он$_i$ сделает доброе дело*[4].

(42b) **Если [каждый бизнесмен]$_i$ вложит деньги в строительство школы, президент даст ему$_i$ награду.*

Обобщение, которое нам удалось сделать, заключается в том, что предложения, в которых возможно связывание, допускают перифразы с выражениями “таким образом”/“тем самым” типа (42).

---

[2] Стоит сразу отметить, что не все носители русского языка считают подобные примеры приемлемыми, однако, относительно большая часть носителей признают их.

[3] Ср. наблюдение в [Татевосов 2002]: «совершенный вид в русском языке не допускает множественных темпоральных референтов описываемой ситуации».

[4] Рецензенты отмечают, что примеры типа (41a) имеют одинаковую степень приемлемости независимо от формы времени и наклонения. В свою очередь, хочу заметить, что опрошенные мной информанты соглашались с тем, что генерические предложения (i) и (ii) звучат значительно хуже, чем соответствующие предложения в будущем времени (iii) или сослагательном наклонении (iv):

(i) *??Если [каждый бизнесмен]$_i$ вкладывал деньги в строительство школы, он$_i$ делал доброе дело.*

(ii) *??Если [каждый бизнесмен]$_i$ вкладывает деньги в строительство школы, он$_i$ делает доброе дело.*

(iii) *Если [каждый бизнесмен]$_i$ вложит деньги в строительство школы, он$_i$ сделает доброе дело.*

(iv) *Если бы[каждый бизнесмен]$_i$ вложил деньги в строительство школы, он$_i$ бы сделал доброе дело.*

Возможно, будет целесообразным провести экспериментальное исследование с большим количеством информантов с целью выяснения того, есть ли контраст между предложениями (i)–(ii), с одной стороны, и (iii)–(iv) с другой.

Такие перифразы невозможны с предложениями, недопускающими связывание, как показывает (43):

(84) Если бизнесмен вкладывает деньги в строительство школы, он тем самым делает доброе дело.

(85) *Если бизнесмен вкладывает деньги в строительство школы, он тем самым получает деньги от государства.

Таким образом, кажется, что для возможности связанной интерпретации необходимо специальное отношение между зависимой и главной клаузами — событие в главной клаузе должно являться естественным/автоматическим результатом события в зависимой клаузе. [5] Более формальное определение условий, влияющих на возможность связывания, остается за рамками настоящей работы.

4. Заключение

Употребление местоимений с кванторным антецедентом в русском языке подчиняется линейному ограничению: квантифицированная ИГ должна линейно предшествовать местоимению, которое она связывает, на одном из этапов синтаксической деривации (ср. формулировку в терминах подъема квантора в [Ivlieva 2009]). То, что это ограничение может накладываться на разных этапах деривации, предсказывает возможность связывания в примерах типа (13–16), где в поверхностной структуре местоимения предшествуют кванторам, а в исходной, наоборот, кванторы предшествуют местоимениям.

Было поставлено под сомнение условие структурного приоритета, предложенное Рейнхарт. Было приведено несколько типов конструкций, в которых это условие не соблюдено, однако, связанная интерпретация доступна.

Были также обнаружены примеры, ставящие под сомнение ограничение, гласящее, что кванторы, находящиеся в зависимых клаузах, не могут связывать местоимения в главных клаузах. Сделана попытка определения условий, при которых возможно связывание местоимения в главном предложении квантором из зависимого. Обобщение заключается в том, что связывание невозможно в предложениях с глаголом СВ в прошедшем времени, в генерических предложениях. Для возможности связывания в предложениях с будущим временем и сослагательным наклонением необходимо специальное отношение между антецедентом и консеквентом. Наличие подобных примеров свидетельствует о том, что существующие на сегодняшний день теории не способны адекватно описать все факты. Нужен анализ, позволяющий объяснить, почему в контекстах, описанных выше, связывание является возможным.

---

[5] Рецензенты предложили следующее ограничение на то, когда подобное связывание возможно: «Представляется, что условием, разрешающим такие высказывания, является не структура событий (как утверждает автор на стр. 11), а статус термовых выражений: если есть возможность трактовать приведенную структуру как выражение с генерическим термом “Тот X, который вкладывает деньги в финансирование спортивной команды, поступает хорошо”, эта группа информантов *(информанты, котороые признают такие примеры грамматичными* — Н. И.) допускает и соответствующие парафразы вида “Если [каждый X]$_i$ вкладывает деньги в финансирование спортивной команды, он$_i$ /она$_i$/они$_i$ поступает хорошо”». В данной формулировке мне не совсем понятно, что рецензенты имеют в виду, когда говорят о «возможности трактовать приведенную структуру как выражение с генерическим термом». Почему, например, мы не можем проинтерпретировать предложение (35) как выражение с генерическим термом, а именно чем парафраза с генерическим термом «Тот преподаватель, который задерживается на работе, не ужинает» или «Если преподаватель задерживается на работе, он не ужинает» хуже парафразы, приведенной выше?

Литература

1. *Казенин К. И.* Структура составляющих vs. линейный порядок (на примере квантора и местоимения) // Труды международной конференции Диалог-2000. М.:2000.
2. *Падучева Е. В.* Высказывание и его соотнесенность с действительностью // М.: Наука, 1985.
3. *Татевосов С. Г.* Семантика составляющих именной группы: кванторные слова // М.: ИМЛИ, 2002.
4. *Bresnan J.* Linear order vs. syntactic rank: Evidence from weak crossover // Proceedings of CLS 30, 1994.
5. *Bresnan J.* Morphology competes with syntax: Explaining typological variation in weak crossover effects // Pilar Barbosa et al. (ed.), Is the Best Good Enough? Optimality and Competition in Syntax. Cambridge, MIT Press: 1998. P. 59–92.
6. *Chomsky N.* Conditions on rules of grammar // Linguistic Analysis 2. 1976. P. 303–351.
7. *Chomsky N.* Minimalist program // Cambridge, MIT Press: 1995.
8. *Huang J.* Reconstruction and the structure of VP: some theoretical consequences // Linguistic Inquiry 24. 1993. P. 103–138.
9. *Ivlieva N.* Constraints on Pronominal Binding in Russian // B. Wiland (ed.) Proceedings of the 2nd Student Conference on Formal Linguistics. Poznan: AMU, 2009.
10. *Langacker R.* On pronominalization and the chain of command // Reibel & Shane (eds.) Modern studies in English. New Jersey: Prentice Hall, 1969. P. 160–186.
11. *May R.* The Grammar of Quantification. MIT dissertation. 1977.
12. *Postal P.* Cross over phenomena // New York: Holt, Rinehart and Winston, 1971.
13. *Reinhart T.* Anaphora and semantic interpretation // Chicago: Chicago University Press, 1983a.
14. *Reinhart T.* Coreference and bound anaphora: a restatement of the anaphora questions // Linguistics and Philosophy 6. 1983b. P. 47–88.
15. *Reuland E. & Avrutin S.* Binding and Beyond: Issues in Backward Anaphora // Branco, McEnery and Mitkov (eds.) Anaphora Processing. John Benjamins Publishing: 2005. P. 139–162.

Прагматика еды: коннотации в русской и немецкой пищевой лексике[1]

Pragmatics of food. Connotations in russian and german nutrition vocabulary

Иомдин Б. Л. (iomdin@ruslang.ru),
Институт русского языка им. В. В. Виноградова РАН

Пиперски А. Ч. (apiperski@gmail.com),
Московский государственный университет им. М. В. Ломоносова

На систематизированном авторами материале русской и немецкой пищевой лексики (в обоих языках было исследовано около 250 понятий) рассматриваются устойчивые ассоциации слов и проблема их описания в рамках теории коннотации. Предлагаются некоторые новые лексикографические решения.

Вводные замечания

В классической работе Ю. Д. Апресяна отмечается, что «теория коннотации будет более надежно обеспечена, когда (и если) будет завершена эмпирическая работа по составлению достаточно полного словаря хотя бы одного языка, содержащего последовательное описание коннотации лексем на основе интуиции хотя бы одного лексикографа» [Апресян 1995: 163]. Не претендуя на попытку решения этой масштабной задачи (которая, впрочем, как хочется надеяться, в отношении русского языка будет в основном выполнена в Активном словаре русского языка, см. [Апресян и др., в печати]), в данной работе мы на материале двух языков (русского и немецкого) рассматриваем один небольшой, но интересный класс лексем, традиционно считающийся «богатым на коннотации», — пищевую лексику. Специальное исследование такого рода, насколько нам известно, не проводилось; даже в весьма подробной монографии [Вайс, Занадворова и др. в печати], где, в частности, проводится сравнительный анализ русской и немецкой лексики, эта тема практически не затрагивается.

1. Регулярные коннотации в пищевой лексике

Как оказывается, и в русской, и в немецкой пищевой лексике самая частая коннотация — оценочная. При этом в русском языке слов с положительной оценкой больше, чем слов с отрицательной оценкой, а в немецком языке — наоборот. В исследованном нами материале, включающем в себя более 250 лексем со значением основных пищевых продуктов и блюд, обнаружилось следующее распределение (разумеется, числа не претендуют на абсолютную точность, но достаточно надежно отражают ситуацию в целом):

	Положительная оценка	Отрицательная оценка	Положительная оценка / отрицательная оценка
Русский	22	15	≈ 1,5
Немецкий	17	27	≈ 0,6

[1] Работа выполнена при финансовой поддержке Программы фундаментальных исследований отделения историко-филологических наук РАН «Генезис и взаимодействие социальных, культурных и языковых общностей» и гранта НШ-4019.2010.6 для поддержки научных исследований, проводимых ведущими научными школами РФ. Использован Национальный корпус русского языка (www.ruscorpora.ru) и корпуса Института немецкого языка (www.ids-mannheim.de/cosmas2/).

1.1. Положительная оценка

Ср. ***бальзам*** *(на душу)* (ср. толкование Е. Э. Бабаевой: ‘то, что вызывает положительные эмоции и делает душевную боль меньше’); *На* ***десерт*** *в «Новостях» всегда рассказывают о чем-то, как это принято говорить, «позитивном»* («Октябрь», 2001); ***калачом*** *не заманишь*; *наши вичи едят одни* ***калачи***; *ни за какие* ***коврижки***; *Потом вложим деньги. И сделаем из ЖЭКа* ***конфетку*** («Известия», 2003); *не жизнь, а* ***малина***; *у нас всё в* ***шоколаде***; ***медоточивый***, ***медом*** *намазано*; *солдатская служба не* ***мед***; *жена у него не* ***сахар***; *все как по* ***маслу***; ***маслом*** *по сердцу*; *как* ***сыр*** *в* ***масле*** *кататься*; *кашу* ***маслом*** *не испортишь*; ***умаслить***; *как* ***огурчик***; ***сливки*** *общества*; *молодец-****огурец***. В немецком: *die kleine Partei ist die* ***Hefe*** *bei der Verwirklichung der Reformen* ‘Эта маленькая партия — это движущая сила (букв. дрожжи) в осуществлении реформ’; *allererste* ***Sahne*** *sein* ‘быть первосортным, букв. быть самыми первыми сливками’; *In der Geschichte ist keine* ***Würze*** ‘Эта история неинтересная, букв. В этой истории нет пряности’; *Deine Idee ist* ***Zucker*** ‘Твоя идея прекрасна, букв. Твоя идея — сахар’.

1.2. Отрицательная оценка

Ср. *остаться <оставить> на* ***бобах***; ***каша*** *в голове <во рту>*; *развесистая* ***клюква***; ***лапшу*** *на уши вешать*; *Терри Уоллису было 19 лет, когда он попал в аварию и стал* ***«овощем»***[2]; *…действительность искусства, состоящая не только из празднеств и торжественных дней, составленная не только из гениев и талантов, а размазанная по поверхности унылой действительности тонким слоем заурядного эстетического* ***повидла*** («Октябрь», 1998); ***уксусное*** *лицо <речи>*; ***хрен*** *редьки не слаще.* Ср. также в немецком: ***Kuchen!*** ‘Не тут-то было!, букв. Пирог!’; ***Wurst/Wurscht/Powidl*** *sein* ‘быть безразличным, букв. быть колбасой/повидлом’; *Wozu der ganze* ***Zimt?*** ‘Для чего вся эта ерунда, букв. корица?’; *jmdn. durch den* ***Kakao*** *ziehen* ‘подшучивать над кем-л., букв. протягивать кого-л. через какао’.

Далее в порядке убывания частотности следуют коннотации, связанные с физическими характеристиками продуктов питания: с их формой, консистенцией и цветом.

1.3. Форма

Ср.: ***банан*** ‘водные сани’, *брюки-****бананы***, *заколка-****банан***; ***блин*** ‘диск от штанги’ (спорт. жарг.); ***ватрушка*** ‘круглый надувной баллон для катания с ледяной горки’; ***колбаса*** ‘тип аэростата’; буфер трамвайного вагона’; *хвост* ***бубликом***; *платье в* ***горошек***; *кони в* ***яблоках***; *нос* ***картошкой***; *пальцы как* ***сардельки*** *<как* ***сосиски****>*; *Ауди-****селедка***; *По-настоящему их обоих звали Николаями, а Клюквой и Килькой прозвали только для различия. У Клюквы были очень красные щеки, а Килька был тощ и вертляв, словно* ***килька*** (Е. Ильина, Четвертая высота); *Забыть Феденьку невозможно. Карикатурно вытянут в длину, как* ***макаронина*** («Дружба народов», 1998). В немецком: ***Birne*** ‘1. груша; 2. электрическая лампочка; 3. (просторечн.) голова’; ***Erbse*** ‘1. горох; 2. (просторечн.) голова’; ***Kartoffel*** ‘1. картофель; 2. (разг.) большие наручные/карманные часы; 3. (разг.) большая дырка (в чулке и т. п.); 4. (разг.) плохой футбольный мяч’; *Atom****pilz*** ‘атомный гриб’; ***Zwiebel****dach* ‘луковичный купол, букв. луковичная крыша’.

1.4. Консистенция

Самое распространенное переносное значение в этой группе как в русском, так и в немецком языке — ‘аморфная смесь с уничтожением отдельных компонентов, иногда до полной неразличимости’. Ср. *Я из тебя* ***котлету*** *<****бифштекс, компот****> сделаю*; *рассыпаться в* ***муку***; *разорвать в* ***лапшу***; *получилась какая-то* ***каша***; *Дорогу развезло: под ногами* ***кисель***; *Читать мне нравится одновременно романов десять. Я оставляю книгу в раскрытом виде и принимаюсь за другую. Через неделю в голове получается полный* ***компот*** (И. Грошек, Легкий завтрак в тени некрополя); *Экспозиции представляют собой невообразимый* ***салат*** *из кхмерской скульптуры, персидских терракот, китайских бронз эпохи Мин, произведений Матисса, Арпа, Бранкузи и современных дюссельдорфских художников* («Вокруг света», 2004); *На нем — целый* ***коктейль*** *чувств: изумление, страх, восторг, отчаяние…* (В. Громов, Компромат для олигарха); *И все это вместе такой конгломерат огромный, такой* ***бульон*** *демократический* (Радио «Свобода»); *Через раскрытые двери было видно, как мимо входа, из одного зала в другой ползет плотная, смятая до состояния* ***пюре*** *людская масса* («Известия», 2003); *Взрывная волна прошлась по помещениям прихотливо: растерев в* ***рагу*** *и щебенку все и всех в полуподвальчике, где был центр связи и информации, она пощадила пост охраны напротив* (А. Лазарчук, Все, способные держать оружие); *С этого момента судьбы троих людей сплетаются в тесный клубок. Их смешивают в* ***коктейль***, *прокручивают в* ***фарш*** («Совершенно секретно», 2003); *В момент отрезвления, точнее,*

---

[2] Интересно, что слово *овощ* в окказиональном значении ‘парализованный человек, утративший способность общаться и воспринимать окружающее’ встречалось и значительно раньше: *Телега медленно движется, вся белая, а я в ней точно* ***овощ****: лежи и молчи, вытянув ноги, да посматривай за знакомыми и считай число зевков у родных, а на подушке незабудки из глины, шныряют прохожие* (В. Хлебников, Мирсконца, 1912). Ср. также англ. *vegetable* ‘овощ’, которое имеет то же переносное значение.

на период отрезвления ты лишаешься человеческих прав, тобой можно крутить, вертеть, подвешивать тебя, сгибать, разгибать. Из тебя можно сделать ***отбивную, шашлык, фарш*** *для кулебяки* (Ю. Азаров, Подозреваемый). Ср. также в немецком: *In RocketKitKongoKit (1986) macht er aus den westlichen Imperialisten in Belgisch-Kongo am Schneidetisch* ***Hackfleisch*** ‘В (фильме) «RocketKitKongoKit» (1986) он на разделочном столе делает фарш из западных империалистов в Бельгийском Конго’ («Die Presse», 1995); *Ich mache* ***Gulasch*** *aus dir!* ‘Я из тебя гуляш сделаю!’; *…ergoss sich der zähe langsame* ***Brei*** *seiner Rede* ‘лилась густая медленная каша его речи’ (H. Fallada, Kleiner Mann — was nun?); *draußen ist eine dicke* ***Soße*** ‘на улице густой туман, букв. густой соус’.

1.5. Цвет

Многие слова имеют стойкие ассоциации с соответствующим цветом; ср. ***молочно****-белый,* ***оливковый, персиковый, апельсиновый, клюквенный, земляничный*** (гораздо реже ***клубничный***), *красный как* ***помидор****, красный как* ***рак*** (сравнение именно с вареным, а не с живым раком!); *rot wie ein* ***Krebs*** ‘красный как рак’, ***milch****weiß* ‘молочно-белый’, ***quitten****gelb* ‘желтый как айва’, ***schokoladen****braun* ‘коричневый, шоколадного цвета’, ***tomaten****rot* ‘красный как помидор’.

Остальные коннотации носят более или менее индивидуальный характер. Ср. следующие примеры: *Станция — под землей, а выше — шесть этажей паркинга. Такой слоеный метро-****бутерброд*** *скоро будет достроен* (Вести-Москва); *…нанесение тонкого слоя аморфного кремния на основу из кремния монокристаллического. Такой гибридный* ***сандвич*** *сулит сочетание приличной эффективности и умеренной цены* (Мембрана); *Так, например, плоскостные молекулы хризотилового асбеста имеют слоистую несимметричную структуру, вследствие чего они сворачиваются в очень тонкую трубочку (своеобразный «****рулет****»)* («Текстиль», 2002) (‘многослойность’); *юмор с* ***горчичкой***; *фельетон с* ***перцем*** (‘острота’); *на* ***десерт*** *<на* ***закуску****>* ‘самое лучшее, появляющееся в конце’; *с* ***изюминкой*** (‘небольшое, но ценное’); *крепкий <твердый>* ***орешек*** (‘сложное’); *это для нас* ***семечки*** (‘простое’); *ни под каким* ***соусом****; Под* ***соусом*** *историко-революционной романтики удалось до некоторой степени легализовать бардовские таланты Владимира Высоцкого, чье бунтарское творчество в ту пору находилось за рамками официальной советской культуры* («Известия», 2001) (‘внешнее, не отражающее сути’); *Всем выступающим дается понять, что они —* ***гарнир, приправа*** *к горячему блюду* (Г. Горин); *В «Итальянцах в России» трюки являются не* ***гарниром****, а мясом, они — содержание произведения* (Э. Рязанов) (примеры из материала М. Я. Гловинской) (‘не основное’); ***чаи*** *распивать <гонять>*, ***чаевничать*** (‘отдых’). Ср. также в немецком: *Manchmal ist die* ***Hefe*** *bloss ein Satz oder ein Wort, das unser Leben nachhaltig verändert* ‘Иногда движущей силой, букв. дрожжами оказывается одно только предложение или слово, которое надолго изменяет нашу жизнь’ («Die Südostschweiz», 2008); *… die wenigstens im Deutschen Sportfernsehen (DSF) gezeigt wurde, allerdings als* ***Konserve*** *zu nachtschlafender Zeit* ‘(чемпионат), который показывали по крайней мере по Спортивному телевидению Германии (DSF), хотя и в записи (букв. как консервы) в ночное время’ («Die Presse», 1997) (‘долго хранящееся’); *eine harte* ***Nuss*** *zu knacken geben* ‘задать трудную задачу, букв. дать расколоть твердый орех’ (‘сложное’); ***Sandwich****mann* ‘человек-бутерброд’ (‘многослойность’); *abwarten und* ***Tee*** *trinken* ‘надо выждать, букв. выжидать и пить чай’ (‘отдых’).

2. Чистые коннотации: есть ли они в русской пищевой лексике?

Как известно, коннотации не являются частью значения слова и логически из него не вытекают. Ср. определение Ю. Д. Апресяна: «коннотациями лексемы мы будем называть несущественные, но устойчивые признаки выражаемого ею понятия, которые воплощают принятую в данном языковом коллективе оценку соответствующего предмета или факта действительности. Они не входят непосредственно в лексическое значение слова и не являются следствиями или выводами из него» [Апресян 1995]. Если применить это определение к приведенным выше примерам, то становится видно, что на самом деле подлинных коннотаций в русской пищевой лексике совсем немного.

Большинство рассмотренных нами дополнительных признаков, связанных с названиями продуктов и блюд, если и не входят в прямое значение лексемы, то непосредственно вытекают из него. Таковы, в частности, признаки формы, цвета, консистенции. Что касается оценок, то большинство из них также вполне объяснимы и системны. Так, потенциальные положительные коннотации имеют все названия сладостей и кондитерских изделий. Некоторые из них устойчивы и способствовали развитию соответствующих переносных значений (ср. *конфетка, малина, пирог*), другие проявляются во фразеологии (ср. *сахар, мед, калач, коврижка, пряник*). Однако более или менее окказиональные переносные употребления с положительной оценкой возможны и для остальных слов этого семантического поля, даже достаточно редких. Ср. *То есть литература, на ваш взгляд, не должна быть сладким* ***пирожным*** *или другой «духовной пищей», а должна быть бритвой, ножом, режущим по живому?* («Знамя», 2003); *Бабушка присылала внуку ласковые от-*

крытки, начинавшиеся «Эдинька, радость, прелесть и ***пончик!****»* (Э. Лимонов, У нас была Великая Эпоха); *Кроме того, обещаны некие интересные* ***плюшки****, подготавливаемые специально для выставки — возможны даже детали Project Natal. Ждем* (Сайт компьютерных игр Maxigame.By); *Ваня вырос не в богатстве, не в* ***марципанах****. И в коммуналке жил* («Собеседник», 2009); *В слоеном пироге шведо-русско-финского гельсингфорсского общества флотские офицеры вкраплены блестящими* ***цукатами*** *в верхний, лучший слой; они — украшение, блеск и вкус* (Л. Соболев, Капитальный ремонт). Любопытно также широко представленное в текстах варьирование фраземы *ни за какие коврижки*: *ни за какие пироги <пряники, ковриги, плюшки, пирожки,…>* (ср. [Беликов 2008] о проблеме вариативности паремий). Положительные коннотации *сладкого* общеизвестны, хотя, по-видимому, в русском языке это явление достаточно новое, а прежде в такой роли в гораздо большей мере выступало *соленое*[3]. *Положительные коннотации масла* (питательное, мягкое, легко поддающееся любому воздействию) и *сливок* (верхняя, лучшая часть молока) тоже объяснимы без труда.

Отрицательная оценка в русском языке значительно реже сопутствует словам с пищевой семантикой. Единственное отмеченное нами системное явление — приведенная выше лексика, связанная с идеей аморфной смеси, обычно содержащая отрицательную оценку. Остальные случаи имеют индивидуальные и иногда не совсем ясные причины. Ср. ***клюква*** (связано с выражением ***развесистая клюква***, ср. СУш); ***бобы*** (связано с гаданием на бобах, ср. СД, СУш, или с дешевизной бобов); ***петрушка*** (связано с названием театральной куклы, ср. Успенский 1962).

3. Чистые коннотации в немецкой пищевой лексике

В немецком языке число чистых коннотаций «пищевых» слов существенно выше, чем в русском. Приведем обнаруженные нами примеры (нельзя не отметить, что большинство из них содержат отрицательную оценку; в частности, очень многие слова имеют значение ‘ерунда, чепуха’): *seinen* ***Senf*** *dazu geben* ‘вмешиваться в разговор, букв. добавлять своей горчицы’, *einen langen* ***Senf*** *machen* ‘тянуть волынку, букв. делать долгую горчицу’; *jmdm.* ***Wurst/Wurscht/Powidl*** *sein* ‘быть безразличным кому-л., букв. быть для кого-л. колбасой/повидлом’; ***Zimt*** ‘1. корица; 2. хлам; ерунда, чепуха’; *danke für* ***Backobst!*** ‘ни при каких условиях!, букв. ‘спасибо за (ненужные) сухофрукты’); ***Kohl*** ‘1. капуста; 2. ерунда, чепуха’; ***Käse*** ‘1. сыр; 2. ерунда, чепуха’: *Hast du nichts Gescheiteres zu tun, als solchen* ***Käse*** *zu schreiben?* ‘Ты не можешь найти себе более разумного занятия, чем писать такую ерунду?’ (J. S. Hohmann, Entstellte Engel); *so ein* ***Quark!*** ‘какая ерунда!, букв. какой творог!’; *das interessiert mich einen* ***Quark*** ‘это меня совершенно не интересует, букв. это интересует меня (как) творог’ (‘ненужное, малоценное’), ***Schinken*** ‘толстая книга, большая (плохая) картина (разг.), букв. ветчина, окорок’, *den* ***Schinken*** *nach der* ***Wurst*** *werfen* ‘рисковать многим ради малого, букв. бросать ветчину ради колбасы’ (‘величина’); *jmdm. ein(en)* ***Bonbon*** *ans Hemd kleben* ‘одурачить кого-л., букв. наклеить кому-л. конфету на рубашку’ (‘глупость’); *Denn kein Führer führt aus dem* ***Salat*** ‘Потому что никакой фюрер не избавит от неприятностей, букв. не выведет из салата’ (B. Brecht, Aufbaulied der F.D.J.); *die* ***Suppe*** *auslöffeln* ‘расхлебывать кашу, букв. расхлебывать суп’ (‘неприятности’); *der fährt eine ganz müde* ***Gurke*** ‘он ездит на старой развалюхе, букв. на усталом огурце’ (‘непригодность’); *für ein Stück* ***Brot*** *<einen* ***Apfel,*** *ein* ***Ei>*** ‘очень дешево, букв. за кусок хлеба <за одно яблоко, за одно яйцо>’ (‘дешевизна, простота’); *Damit ist jetzt* ***Essig*** ‘С этим дело покончено, букв. с этим уксус’ (‘несовременность’); *jmdn. auf die* ***Nudel*** *schieben* ‘обманывать кого-л., букв. толкать на вермишель’: *Wir lassen uns nicht länger auf die* ***Nudel*** *schieben* ‘Мы не позволим нас больше обманывать’ («Berliner Zeitung», 2001) (‘обман’); ***Grütze*** ‘1. крупа, каша; 2. ум’: *ich habe mehr* ***Grütze*** *im Kopf als ihr alle zusammen* ‘у меня в голове больше мозгов (букв. больше крупы), чем у вас всех вместе взятых’ (H. Jaeger, Das Freudenhaus).

4. Источники коннотаций в пищевой лексике

Как нам кажется, можно выделить три важнейших источника коннотаций в пищевой лексике.

4.1. Естественные признаки

Переносные значения могут непосредственно вытекать из основных, т. е. следовать из наиболее характерных свойств денотата, например его цвета, формы, вкуса, консистенции. Так, вполне естественно, что *помидор* ассоциируется со значением

---

[3] Ср. «Несмотря на семантические расхождения, все перечисленные фонетические варианты корней **sol-//*sold-* в русском языке обладают значением ‘вкусный; быть/делать вкусным’ (ср. арх. *подсоли́ть* ‘приправить для вкуса, сдобрить’ — «Чаек-то ваш подсолю молочком», пск. *засолоди́ть* ‘приправить’, разг. *сладкий* ‘вкусный’). В противоположность «безвкусному» пресному и пустому соленая и сладкая пища с полным правом может называться вкусной, ср. поговорку *Щи капустою пригожи, а солью укусны*, загадку *Что на свете всех вкуснее*? (соль). <…> Однако если соль — это обязательная составляющая пищи <…>, то сладкий вкус, хотя и воспринимается как «положительный», является непривычным для традиционной русской культуры (своего рода излишеством)» [Пьянкова 2008: 11].

‘красный’, *колбаса* — со значением ‘продолговатый’, *мед* — со значением ‘приятный’, а *коктейль* — со значением ‘смесь’.

4.2. Прагматическая оценка

Значительную роль играют внешние условия употребления продукта, т. е. то, предназначен ли он для одного человека или для многих, его распространенность, обыденность, дешевизна или, наоборот, доступность только богатым, а также некоторые другие факторы. Так, в русском языке *пирог* связывается с идеей деления (поскольку пирог обычно едят несколько человек4), а *репа* как один из самых распространенных овощей — с идеей простоты (*проще пареной репы*). В немецком как нечто дешевое, незначительное концептуализируются *Apfel* ‘яблоко’, *Bohne* ‘боб’, *Ei* ‘яйцо’. В свою очередь, и в русском, и в немецком языке как нечто ценное воспринимается *шампанское* (ср. *кто не рискует, тот не пьет шампанского*; *Sekt oder Selters* ‘всё или ничего, букв. шампанское или сельтерская’).

4.3. Созвучие

Коннотации могут возникать и по созвучию с другими словами. Например, немецкое слово *Kakao* имеет коннотацию ‘насмешки; нечто неприятное’ (*jmdn. durch den Kakao ziehen* ‘подшучивать над кем-л.’), вероятно, являясь при этом эвфемистической заменой для слова *Kacke* ‘экскременты’ [Duden]. Русское слово *батоны* имеет жаргонное значение ‘ягодицы’. Эту связь нельзя объяснить формой батона; можно предполагать, что это перенос значения со слова *булки* (ср. в песне М. Гребенщикова: *Твои батоны, они же булки*), но, с другой стороны, можно думать, что употребление слова *батоны* в этом значении вызвано сходством с англ. *butt*, *buttocks* ‘ягодицы’. Разумеется, если в первом и во втором случае мы имеем дело с некоторыми явлениями, которые особенно типичны для названий еды, то в случае с возникновением коннотации по созвучию пищевая лексика не обнаруживает никакой специфики.

Если рассматривать коннотации в диахроническом аспекте, то обнаруживается, что коннотации первой группы (обусловленные характерными свойствами объекта) по мере развития языка обычно продолжают оставаться прозрачными и понятными (т. е. продолжают функционировать как регулярные коннотации). То же самое нельзя сказать о коннотациях двух других групп. Особенно характерна утрата мотивированности для коннотаций группы 2: например, в наши дни носителям немецкого языка едва ли понятно, почему *Essig* ‘уксус’ связывается с чем-то устаревшим (это объясняется тем, что слишком долго бродившее вино превращается в уксус5). Происхождение коннотаций группы 3 тоже может затемняться: носители современного немецкого языка едва ли осознают источник переносного значения слова *Kakao*.

Впрочем, и коннотации первой группы могут оказаться непрозрачными и претерпевать изменения. В этом отношении интересна история слова *банан*, которое еще недавно в школьном жаргоне означало оценку «2» (ср. *схлопотать банан по физике*). Это значение было перенесено с оценки «1», по форме напоминающей банан, но практически переставшей применяться (см., например, Елистратов 2000). В современном же компьютерном сленге *банан* означает ‘временное или постоянное лишение возможности посылать сообщения (на форумах и т. п.) за некооперативное поведение’ (поиск на форумах *схлопотать банан* дает практически исключительно такие употребления), что поддерживается созвучием с англ. *ban* ‘то же’ (ср. сленговое *забанить*).

5. Коннотации пищевой лексики в словаре: проблемы описания

Представляется очевидным, что «чистые» коннотации должны включаться в словарную статью соответствующей лексемы. Это особенно важно в случае индивидуальных коннотаций, которые часто обнаруживают лингвоспецифичность (ср. положительную коннотацию *огурца* в русском и отрицательную в немецком). Более интересен вопрос о том, в каких случаях следует считать устойчивые ассоциации, сопутствующие некоторым словам, их коннотациями, а в каких — частью значения.

Рассмотрим в качестве примера лексемы *хлеб* и *соль*, которые все толковые словари русского языка определяют примерно так: ‘пищевой продукт, выпекаемый из муки’ и ‘белое кристаллическое вещество с острым вкусом, употребляемое как приправа’, соответственно. Очевидным образом, эти толкования не являются не только исчерпывающими, но даже и дифференциальными: под первое подходят все мучные изделия, от *пирога* до *сушки*, второе из-за неопределенности прилагательного *острый* (тол-

---

[4] По-видимому, раньше в русском языке слово *пирог* могло обозначать и большое изделие, предназначенное для нескольких человек, и маленькое, рассчитанное на одного — то, что теперь обычно называется словом *пирожок*. Ср., с одной стороны, *Здорово, Дашенька! Да вот принес пирог твоим господам. Ведь ты знаешь, барин мой звал ваших сюда на завтрак* (И. А. Крылов, Пирог) и, с другой стороны, *Сначала горячее подадут, как следует, с пирогами, да только уж пироги с наперсток; возьмешь в рот вдруг штук шесть, хочешь пожевать, смотришь — уж там их и нет, и растаяли…* (И. А. Гончаров, Обыкновенная история).

[5] Ср. схожие наблюдения об отрицательных коннотациях уксуса на итальянском материале в [Бушуева 2005: 26].

куемого в тех же словарях как ‘с большим количеством соли, пряностей, специй, раздражительный, едкий, пряный, сильно действующий на вкус или обоняние’) может описывать и лимонную кислоту, и глютамат натрия (отвлекаемся здесь от других проблем — в частности, той, что *мука* в тех же словарях толкуется как размолотые хлебные зерна, что приводит к кругу). Тот факт, что *хлеб* и *соль* имеют в русском языке прочные ассоциации ‘основного, необходимого’ (хлеб — основная пища, соль — необходимая составляющая любой трапезы), подтверждается богатейшим материалом, многократно отмечался и не требует дополнительных комментариев. По всем критериям эти ассоциации безусловно являются коннотациями лексем *хлеб* и *соль*. Однако такое решение явным образом противоречит основному свойству коннотаций, поскольку те по определению воплощают несущественные признаки понятия, выражаемого лексемой [Апресян 1995: 159]. Кажется более разумным включение этих важнейших для представления о хлебе и соли в русской языковой картине мира признаков в толкование лексем. Это решение, в частности, устранило бы некоторые недостатки приведенных толкований традиционных словарей, поскольку сушку нельзя считать основным продуктом питания, а глютамат натрия — необходимой составляющей любой трапезы.

Интересное свойство коннотаций, отмеченное Ю. Д. Апресяном, состоит в том, что «коннотируют обычно родовые слова (*ветер*, но не *суховей*, *свежак*), достаточно употребительные (*резать*, но не *нарезать*, *стрелять*, но не *палить*), не являющиеся терминами (*ветер*, но не *бора*, *бриз* и т. п.)» [Апресян 1995: 173]. Изучение и сравнение коннотаций в классах лексем с близким значением позволит лучше понять внутреннюю организацию лексики в этих классах. В работе [Иомдин 2009] было выделено несколько критериев определения доминанты в группе слов с близким значением, относящихся к бытовой лексике. Еще одним из таких критериев может служить наличие и объем коннотаций. Так, в рассмотренном материале из всех кондитерских изделий наибольшим числом коннотаций обладает *пирог*, и этот факт — дополнительное свидетельство в пользу того, что именно *пирог* (а не *торт*, *кекс*, *шарлотку* и др.) следует признать доминантой соответствующей группы слов.

Вместо заключения

Несмотря на хорошую разработанность темы коннотаций в современной лингвистической литературе, представляется, что она требует дальнейшего изучения и проверки на большом фактическом материале. Результаты такого исследования необходимы в современной лексикографической практике, как толковой, так и двуязычной.

Литература

1. *Апресян* 1995 — Апресян Ю. Д. Коннотации как часть прагматики слова // Избранные труды. В 2-х тт. М.: Языки русской культуры, 1995.
2. *Апресян* и др. в печати — Апресян Ю. Д., Апресян В. Ю., Бабаева Е. Э., Богуславская О. Ю., Иомдин Б. Л., Крылова Т. В., Левонтина И. Б., Санников А. В., Урысон Е. В. Проспект активного словаря русского языка. Под общим руководством акад. Ю. Д. Апресяна. М.: Языки славянских культур, в печати.
3. *Беликов* 2008 — Беликов В. И. Паремии как объект лексикографии // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 4–8 июня 2008 г.). Вып. 7 (14). М.: РГГУ, 2008. С. 45–49.
4. *Бушуева* 2005 — Бушуева В. В. Функционирование базовых лексем кулинарной сферы в современном итальянском языке // Проблемы и методы современной лингвистики. Вып. 1. М.: Институт языкознания РАН, 2005. С. 19–27.
5. *Вайс, Занадворова* и др. в печати — Вайс Д., Занадворова А. В., Иссерс О. С., Китайгородская М. В., Ратмайр Р., Розанова Н. Н., Хоффманн Э. Еда по-русски в зеркале языка. М., в печати.
6. *Елистратов* 2000 — Елистратов В. С. Словарь русского арго: Материалы 1980–1990 гг.: Около 9 000 слов, 3 000 идиоматических выражений. М.: Русские словари, 2000.
7. *Иомдин* 2009 — Иомдин Б. Л. Терминология быта. Поиски нормы // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2009» (Бекасово, 27–31 мая 2009 г.). Вып. 8 (15). М.: РГГУ, 2009, с. 127–135.
8. *Пьянкова* 2008 — Пьянкова К. В. Лексика, обозначающая категориальные признаки пищи, в русской языковой традиции: этнолингвистический аспект. Авт. дисс. … к. ф. н. Екатеринбург, 2008.
9. *СУш* — Толковый словарь русского языка / Под ред. Д. Н. Ушакова. М.: Гос. ин-т «Сов. энцикл.»; ОГИЗ; Гос. изд-во иностр. и нац. словарей, 1934–1940.
10. *Успенский* 1962 — Успенский Л. В. Слово о словах. Л.: Лениздат, 1962.
11. *Duden* — Das große Wörterbuch der deutschen Sprache. 3. Aufl. Mannheim: Bibliographisches Institut, 1999.

Сравнительный анализ статистических алгоритмов синтаксического анализа на основе деревьев зависимостей

A comparative analysis of machine learning dependency tree-based parsing algorithms

Казенников А. О. (kzn@iitp.ru)

ИППИ им. А. А. Харкевича РАН

В работе сравниваются различные подходы к построению синтаксической структуры на основе деревьев зависимостей. Рассматриваются два статистических подхода построения деревьев зависимостей: как задача минимальных остовных деревьев, и как задача разбора с автоматом с очередью. Оба подхода в качестве базового алгоритма используют SVM. Кроме того, производится сравнение эффективности этих подходов относительно системы на основе правил ЭТАП-3.

1. Введение

Традиционно сложная задача обработки текста на естественном языке разбивается на несколько уровней. Обычно, анализ на каждом уровне производится независимо от остальных. В частности, такими отдельными задачами являются морфологический анализ, снятие частеречной омонимии, построение синтаксической и семантической структуры. Поэтому, естественно предполагать, что при таком подходе в качестве исходных данных для задачи синтаксического анализа предстают слова с однозначно определенной частью речи.

Однако это не единственно возможный подход. Например, система ЭТАП-3 [1] проектировалась как лингвистический процессор. Его задачей является анализ текста, начиная с самого первого уровня. При разработке системы ЭТАП-3 было принято решение о том, что омонимия может быть разрешена в ходе синтаксического анализа.

В настоящей работе синтаксический анализ рассматривается как задача построения дерева зависимостей предложения при условии снятой частеречной омонимии.

2. Постановка задачи

Формально задача синтаксического анализа формулируется следующим образом. Дано предложение $S = \{w_i\}, i \in \{1 \dots n\}$, где w_i — i-слово предложения, n — число слов в предложении. В начало предложения добавляется фиктивное слово w_0, которое обозначает вершину синтаксической структуры предложения. Необходимо построить ориентированное дерево $G = (V, A)$, где V — вершины (слова), A — дуги (синтаксические связи). В дереве должна быть одна вершина и не должно быть циклов. Поскольку основным элементом анализа является связь, то необходимо определить несколько ее характеристик:

- хозяин связи — слово, из которого связь выходит,
- слуга — слово, в которое связь приходит,
- левое и правое слова связи —слова связи относительно их порядка в предложении,
- направление связи — в какую сторону (влево или вправо) от хозяина идет связь,
- имя связи — имя синтаксического отношения, связывающего слова между собой.

Эти характеристики являются «внутренними» — они относятся только к связи. Однако есть и очень важная «внешняя» характеристика связи, которая относится к структуре в целом и существенно влияет на алгоритмы синтаксического анализа: это проективность.

Неформально проективность связи можно определить графически — если построить дерево синтаксической структуры, то проективные связи между собой не пересекаются. Если какие-то две связи пересекаются между собой, то одна из них непроективна (см. рис 1). Более формально свойство проективности можно определить следующим образом — из хозяина связи доступны (по связям) все слова, которые эта связь перекрывает.

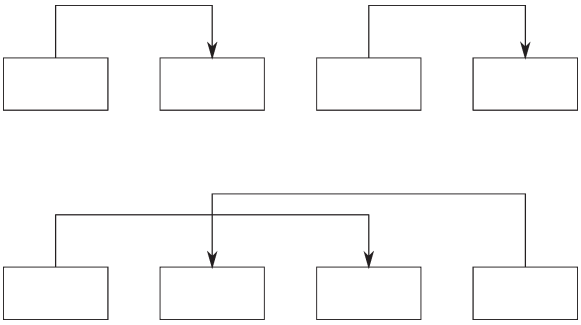


Рис. 1. Проективные и непроективные связи

3. Обзор существующих подходов

В работе рассматривалось несколько подходов к задаче синтаксического анализа:

- Облегченная система ЭТАП,
- Алгоритмы на основе максимальных остовных деревьев,
- Алгоритмы на основе системы переходов.

Система ЭТАП-3 является системой на основе правил. Она в значительной степени отталкивается от лингвистической теории «Смысл ↔ Текст» [2]. Правила для системы пишутся экспертами-лингвистами. На первом этапе синтаксического анализа строится матрица потенциальных связей в предложении, затем из этих связей формируется дерево. Часто существует возможность построения нескольких вариантов структуры для одного предложения. В таком случае по умолчанию выбирается первая построенная структура. Упрощенно говоря, задачей лингвиста является составление таких правил, при которых первое построенное дерево было бы оптимальным с лингвистической точки зрения. Таким образом, лингвистическая модель задается в явном виде с помощью правил. Основным недостатком такого подхода является необходимость больших трудозатрат для построения качественной системы (система ЭТАП-3 разрабатывается более 20 лет).

Принципиально другим способом представления модели языка является неявное представление в виде большого размеченного корпуса. Для практического применения такой модели используется подход на основе машинного обучения [3,4]. Тогда структура строится на основе закономерностей, выведенных алгоритмом из корпуса. Существенным недостатком этого подхода является сложность лингвистической интерпретации полученной модели, а так же необходимость в достаточно большом корпусе.

Важной особенностью систем на основе машинного обучения является полная зависимость от качества решения поставленной задачи машинного обучения. Т. е, если задача машинного обучения решена плохо, то соответственно будут плохими и результаты работы такой системы, независимо от используемого подхода.

В настоящем разделе представлен обзор подходов по их схеме работы, а в следующем разделе представлены подробная постановка и решение задачи машинного обучения.

В работах [3,4] был проведен сравнительный анализ статистических алгоритмов и было показано, что наиболее эффективными подходами является представление синтаксического анализа как задачи выделения максимального остовного дерева, а так же представление его как задачи поиска оптимальной последовательности действий.

Подход на основе максимальных остовных деревьев [3] рассматривает задачу синтаксического анализа как задачу нахождения максимального остовного дерева(MST) на графе возможных связей. Предполагается существование функции оценки связи $s(i,j) = \mathbf{w} \cdot \mathbf{f}(i,j)$. где $\mathbf{f}(i,j)$ — признаки, на основе которых принимается решение о проведении связи, w — модель, полученная с помощью машинного обучения.

Алгоритм выбирает такое дерево, сумма оценок связей которого будет максимальна:

$$s(\mathbf{x},\mathbf{y}) = \sum_{(i,j)\in\mathbf{y}} s(i,j)$$

$$\max_y s(\mathbf{x},\mathbf{y})$$

Задача машинного обучения заключается в получении такой функции оценки связи, которая бы позволяла построить правильную структуру, для наибольшего числа предложений из корпуса

Алгоритм построения дерева на основе функции оценки сильно зависит от необходимости построения непроективных связей. В классическом случае, когда *f(i, j)* зависит только от характеристик оцениваемой связи, алгоритм с возможностью построения непроективных связей проще, чем тот, который строит только проективные связи. Однако допущение непроективных связей сильно ограничивает возможность использования дополнительных параметров связи.

Другой подход — подход на основе системы переходов(TS) [4]. Парсер на основе такого подхода состоит из 3 компонентов:

1. Конфигурации — состояния процесса разбора в каждый конкретный момент.
2. Действия, изменяющие конфигурацию.
3. Начальное и конечное состояние конфигурации.

Задача синтаксического анализа сводится к поиску цепочки действий, которая бы переводила начальную конфигурацию в конечную. Для выбора каждого следующего действия используется оракул. Его задача — на основе текущей конфигурации выбрать следующее действие. Задача машинного обучения состоит в моделировании оракула. Параметры парсера очень сильно зависят от выбора систе-

мы действий и выбора параметров конфигурации, на основе которых принимается решение о следующем действии.

4. Обзор реализации подходов

Хотя система ЭТАП-3 разрабатывалась как система машинного перевода, ее архитектура позволяет работать в качестве синтаксического анализатора. В таком случае обработку входного предложения можно с некоторой степенью условности разделить на два этапа: морфологический и синтаксический. Важной особенностью системы ЭТАП-3 является взаимосвязь между модулями, и на вход синтаксического модуля подаются слова с неразрешенной частеречной омонимией. Поэтому, для сравнения системы ЭТАП с другими подходами ее надо уравнять в возможностях с другими системами. В проведенных экспериментах морфологический компонент не участвовал, а входными данными являлись слова со снятой частеречной омонимией.

Для подхода на основе минимальных остовных деревьев необходимо сформулировать принцип построения функции оценки связи. Фактически, задача построения такой функции есть задача ранжирования — необходимо, чтобы эталонная связь (на этапе обучения) получала большую оценку, чем остальные потенциальные связи:

$$s(i,j) > s(k,j), \forall k \neq i$$

$$s(i,j) > s(i,m), \forall m \neq j$$

Для применения такого алгоритма необходимо определить правила для построения рангов. Для деревьев зависимостей их можно определить на основе следующих фактов:

- Для каждого слова правильна только одна входящая связь
- Для данного слова потенциальными хозяевами могут могут быть все остальные слова предложения и вершина.

Другим важным вопросом является выбор модели признаков. Модель признаков определяет преобразование информации о связи и ее участниках в числовой вектор — вектор признаков. Классический вариант алгоритмов максимальный остовных деревьев предполагает независимость ребер графа. Это означает, что каждая синтаксическая связь проводится независимо от уже существующих.

Для подхода на основе системы переходов необходимо определить структуру конфигурации и набор возможных действий над конфигурацией. В работе [4] представлены несколько конфигураций и систем действий на их основе, однако наиболее интересной является конфигурация на основе списков, поскольку с помощью нее возможно построение непроективных структур (остальные системы переходов могут построить только строго проективные структуры).

Конфигурация этой системы состоит из трех списков. Под списком подразумевается линейная структура данных, обладающая вершиной, обозначаемой как head|list, и определенной операцией склейки l_1+l_2. Эти списки можно интерпретировать следующим образом. В списке b хранятся необработанные слова предложения, а вершина списка является потенциальным правым словом связи. Этот список упорядочен в соответствии с порядком следования слов в предложении. В списке l_1 хранятся возможные левые слова связи. Список l_1 расположен в обратном порядке относительно порядка слов в предложении. Список l_2 формируется из слов в промежутке между l_1 и b, порядок слов в нем так же противоположен порядку слов в предложении. Построение связи в какой-либо конфигурации возможно только между вершиной b и вершиной l_1.

При инициализации конфигурации в l_1 помещается фиктивное слово вершины предложения, а в b — слова предложения, l_2 — пуст. Признаком конечного состояния системы(конца разбора) является опустошение списка b.

Система состоит из четырех действий:

1. Left-Arc — проведение левой связи

   $(\lambda_1|i, \lambda_2, j|\beta) \Rightarrow (\lambda_1, i|\lambda_2, j|\beta)$

   связать (j,i)

2. Right-Arc — проведение правой связи

   $(\lambda_1|i, \lambda_2, j|\beta) \Rightarrow (\lambda_1, i|\lambda_2, j|\beta)$

   связать (i,j)

3. No-Arc — пропуск текущего левого слова

   $(\lambda_1|i, \lambda_2, j|\beta) \Rightarrow (\lambda_1, i|\lambda_2, j|\beta)$

4. Shift — пропуск текущего правого слова

   $(\lambda_1, \lambda_2, i|\beta) \Rightarrow (\lambda_1 + \lambda_2|i, [], \beta)$

В работе [4] показано, что сложность разбора на основе такой системы действий равна O(n^2). Можно, однако, показать, что в среднем требуется количество действий, линейно зависящее от числа слов в предложении и средней длины связи. Из определений действий следует, что только операция Shift изменяет список b. Следовательно, для разбора предложения потребуется n действий Shift. В предложении всегда n связей(у каждого слова должен быть хозяин), следовательно, для разбора необходимо n действий Left-Arc или Right-Arc. После операции Shift список l_2 пуст, а потенциальная длина связи равна единице. Каждое действие Left-Arc, Right-Arc

или No-Arc увеличивает длину потенциальной связи на единицу, следовательно, максимальное число действий No-Arc пропорционально средней длине связи предложения. Таким образом, сложность алгоритма такой системы переходов O(n+k*n), где k — средняя длина связи в предложении.

Для создания модели оракула необходим алгоритм преобразования эталонной структуры в последовательность действий. Его можно построить на основе аналогичных соображений. Для преобразования эталонной структуры необходимо:

1. Отсортировать эталонные связи:
   1. По правому слову.
   2. По длине связи.
2. Два индекса: индекс вершины b, индекс вершины l_1.
3. Для каждой эталонной связи:
   1. Пока индекс вершины b не равен индексу правого слова, выполнять Shift.
   2. Пока индекс вершины l_1 не равен индексу левого слова, выполнять No-Arc.
   3. Провести эталонную связь с помощью действий Right-Arc, и Left-Arc
4. Добавить n действий Shift, где n — число оставшихся слов в b.

Обучение производится на основе пар текущая конфигурация → действие. Такая задача является задачей разбиения на несколько классов. В качестве алгоритма обучения используется алгоритм на базе SVM[6].

Одним из существенных отличий описываемой экспериментальной системы от представленной в [4] является модель признаков, извлекаемая из конфигурации. В качестве базовых признаков используются:

1. b[0] — потенциальный участник связи
2. b[1,2,3] — контекст справа
3. l_1[0] — потенциальный участник связи
4. l_1[1,2] — контекст слева
5. hd(l_1[0]) — хозяин l_1[0]
6. ld(l_1[0]) — левое зависимое слово l_1[0]
7. rd(l_1[0]) — правое зависимое слово l_1[0]
8. l_2[0,n] — контекст между потенциальными участниками связи

Кроме того, были добавлены дополнительные признаки, которые могли бы повлиять на качество разбора.

1. ngc(b[0]) — число-род-падеж b[0]
2. ngc(l_1[0]) — число-род-падеж l_1[0]
3. {ngc(b[0]), ngc(l_1[0])} кортеж из b[0] и l_1[0]
4. interior — полный контекст между b[0] и l_1[0]

Кроме того, в модель была неявно включена длина связи. В дополнение к простым признакам связи добавлялись кортежи этих признаков и длины связи.

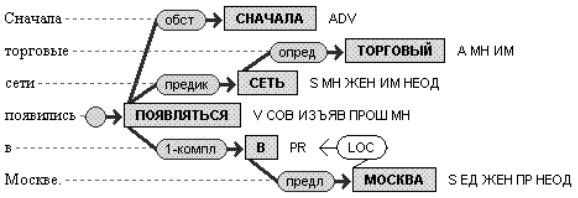


Рис 2. Синтаксическая структура предложения «Сначала торговые сети появились в Москве»

Таблица 1. Иллюстрация разбора с помощью системы действий

Начальная конфигурация	Действие	Построенные связи
[ROOT] [] [Сначала, торговые, сети, появились, в, Москве]	Shift	[]
[ROOT, Сначала] [] [торговые, сети, появились, в, Москве]	Shift	[]
[ROOT, Сначала, торговые] [] [сети, появились, в, Москве]	LeftArc(ОПРЕД)	[$(3\rightarrow2)_{\text{ОПРЕД}}$]
[ROOT, Сначала] [торговые] [сети, появились, в, Москве]	Shift	[$(3\rightarrow2)_{\text{ОПРЕД}}$]
[ROOT, Сначала, торговые, сети] [] [появились, в, Москве]	LeftArc(ПРЕДИК)	[$(3\rightarrow2)_{\text{ОПРЕД}}$,$(4\rightarrow3)_{\text{ПРЕДИК}}$]
[ROOT, Сначала, торговые] [сети] [появились, в, Москве]	NoArc	[$(3\rightarrow2)_{\text{ОПРЕД}}$,$(4\rightarrow3)_{\text{ПРЕДИК}}$]
[ROOT, Сначала] [торговые, сети] [появились, в, Москве]	LeftArc(ОБСТ)	[$(3\rightarrow2)_{\text{ОПРЕД}}$,$(4\rightarrow3)_{\text{ПРЕДИК}}$, $(4\rightarrow1)_{\text{ОБСТ}}$]

Начальная конфигурация	Действие	Построенные связи
[ROOT] [Сначала, торговые, сети] [появились, в, Москве]	RightArc(ROOT)	$[(3\rightarrow2)_{\text{ОПРЕД}},(4\rightarrow3)_{\text{ПРЕДИК}},$ $(4\rightarrow1)_{\text{ОБСТ}},(0\rightarrow4)_{\text{ROOT}}]$
[] [ROOT, Сначала, торговые, сети] [появились, в, Москве]	Shift	$[(3\rightarrow2)_{\text{ОПРЕД}},(4\rightarrow3)_{\text{ПРЕДИК}},$ $(4\rightarrow1)_{\text{ОБСТ}},(0\rightarrow4)_{\text{ROOT}}]$
[ROOT, Сначала, торговые, сети, появились] [] [в, Москве]	RightArc(1-КОМПЛ)	$[(3\rightarrow2)_{\text{ОПРЕД}},(4\rightarrow3)_{\text{ПРЕ-ДИК}},(4\rightarrow1)_{\text{ОБСТ}},(0\rightarrow4)_{\text{ROOT}},$ $(4\rightarrow5)_{\text{1-КОМПЛ}}]$
[ROOT, Сначала, торговые, сети] [появились] [в, Москве]	Shift	$[(3\rightarrow2)_{\text{ОПРЕД}},(4\rightarrow3)_{\text{ПРЕ-ДИК}},(4\rightarrow1)_{\text{ОБСТ}},(0\rightarrow4)_{\text{ROOT}},$ $(4\rightarrow5)_{\text{1-КОМПЛ}}]$
[ROOT, Сначала, торговые, сети, появились, в] [] [Москве]	RightArc(ПРЕДЛ)	$[(3\rightarrow2)_{\text{ОПРЕД}},(4\rightarrow3)_{\text{ПРЕ-ДИК}},(4\rightarrow1)_{\text{ОБСТ}},(0\rightarrow4)_{\text{ROOT}},$ $(4\rightarrow5)_{\text{1-КОМПЛ}},(5\rightarrow6)_{\text{ПРЕДЛ}}]$
[ROOT, Сначала, торговые, сети, появились] [в] [Москве]	Shift	$[(3\rightarrow2)_{\text{ОПРЕД}},(4\rightarrow3)_{\text{ПРЕ-ДИК}},(4\rightarrow1)_{\text{ОБСТ}},(0\rightarrow4)_{\text{ROOT}},$ $(4\rightarrow5)_{\text{1-КОМПЛ}},(5\rightarrow6)_{\text{ПРЕДЛ}}]$
[ROOT, Сначала, торговые, сети, появились, в, Москве] [] []		$[(3\rightarrow2)_{\text{ОПРЕД}},(4\rightarrow3)_{\text{ПРЕ-ДИК}},(4\rightarrow1)_{\text{ОБСТ}},(0\rightarrow4)_{\text{ROOT}},$ $(4\rightarrow5)_{\text{1-КОМПЛ}},(5\rightarrow6)_{\text{ПРЕДЛ}}]$

5. Эксперименты

Для оценки рассматриваемых подходов были проведены эксперименты. В качестве материала использовался корпус СинТагРус [8]. На настоящий момент в корпусе около 40 тысяч предложений, около 600 тысяч слов. В текстах используется порядка 32 тысяч лемм.

Эксперименты проводились следующим образом. Корпус делился на две части — часть для обучения, а часть для оценки полученной модели. В результате эксперимента предполагалось оценить эффективность каждого из подходов на основе машинного обучения по отношению к системе ЭТАП. Поскольку алгоритмы построения минимального остовного дерева сильно зависят от возможности построения непроективных связей, то системы сравнивались по промежуточному параметру — количеству правильно построенных связей.

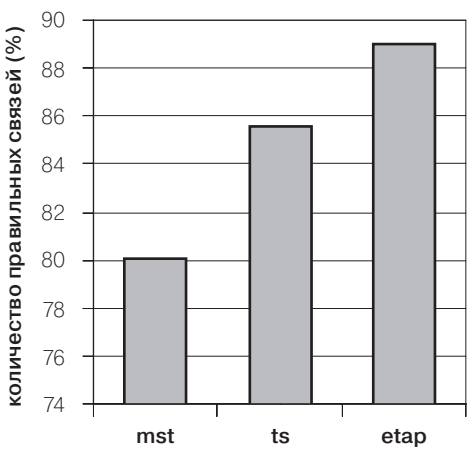


Рис. 3. Процент правильных связей

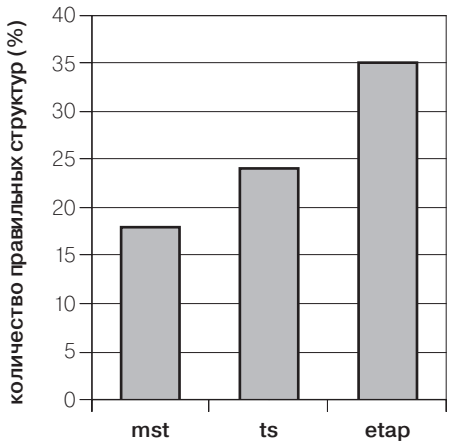


Рис. 4. Процент правильных структур

Для обучения использовалось 10 тысяч предложений, для тестирования — 5 тысяч. Для подхода на основе максимальных остовных деревьев строилось две системы принятия решений: для оценки возможности связи(проведение связи без имени), и восстановление имени связи. Такое разделение было необходимо из-за вычислительной сложности подхода.

На рис. 3 представлены данные по точности построения связей при каждом из рассмотренных подходов(mst — подход на основе максимальных остовных деревьев, ts — на основе системы переходов, etap — система ЭТАП-3). Из этих данных следует, что в абсолютных величинах разница между подходом на основе системы переходов и эталонной системой ЭТАП-3 не такая большая(около 4 %). Однако, такая разница дает более 10 % разницы в количестве построенных структур.

6. Результаты и перспективы

В работе приведена экспериментальная реализация подхода к синтаксическому анализу на основе системы переходов. Несмотря на то, что система на основе машинного обучения не показала лучшие результаты, она является довольно конкурентоспособной. В настоящей работе использовался линейный вариант SVM, в результате чего возник небольшой проигрыш относительно результатов, представленных в [4]. Однако, этот факт компенсируется значительным увеличением скорости работы парсера.

В настоящей работе модель признаков была богаче, чем в [4], поскольку представленная система ориентирована на использование совместно с корпусом СинТагРус, где представлены более богатые лингвистические характеристики.

Поэтому представляется перспективным продолжить работы в области создания статистического парсера, одновременно с этим изучая возможности создания гибридной системы.

Литература

1. *Апресян Ю. Д., Богуславский И. М., Иомдин Л. Л.* и др. Лингвистическое обеспечение системы ЭТАП-2. М., Наука, 1992
2. *Мельчук И. А.* Опыт теории лингвистических моделей «Смысл ↔ Текст». М., Наука,1974.
3. *McDonald R., Crammer K., Pereira F.* Spanning Tree Methods for Discriminative Training of Dependency Parsers. // ACL-06
4. *Nivre J.* Algorithms for Deterministic Incremental Dependency Parsing // Comp. Linguistics Vol 34, No 4.
5. *McDonald R., Satta G.* On the Complexity of Non-Projective Data-Driven Dependency Parsing. // IWPT-2007
6. *Crammer K., Singer Y.* Ultraconservative Algorithms for Multiclass Problems. // JMLR '01
7. *Апресян Ю. Д.* Идеи и методы современной структурной лингвистики. // М. Просвещение, 1966.
8. *Апресян Ю. Д., Богуславский И. М., Иомдин Б. Л., Иомдин Л. Л., Санников А. В., Санников В. З., Сизов В. Г., Цинман Л. Л.* Синтаксически и семантически аннотированный корпус русского языка: современное состояние и перспективы // Национальный корпус русского языка 2003–2005 г. (результаты и перспективы). М: «Индрик», 2005. С. 193–214

База данных по многозначным качественным прилагательным и наречиям русского языка

The database on russian polysemous adjectives and adverbs

Карпова О. С. (o_k_@inbox.ru), РГГУ

Резникова Т. И. (tanja.reznikova@gmail.com), ВИНИТИ РАН

Архангельский Т. А. (timarkh@gmail.com), МГУ им. М. В. Ломоносова

Кюсева М. В. (pegas@nm.ru), МГУ им. М. В. Ломоносова

Рахилина Е. В. (rakhilina@gmail.com), ИРЯ им. В. В. Виноградова РАН

Рыжова Д. А. (daska1990R@yandex.ru), МГУ им. М. В. Ломоносова

Тагабилева М. Г. (geratagabileva@gmail.com), МГУ им. М. В. Ломоносова

В докладе представляется База данных по многозначным качественным прилагательным и соответствующим им наречиям, разработанная на материале Национального корпуса русского языка. Описываются структура Базы, принципы ее заполнения и поисковые возможности, а также некоторые теоретические результаты анализа полисемии в русской признаковой лексике.

1. Постановка задачи

Настоящая работа представляет результаты проекта, посвященного созданию Базы данных по многозначным качественным прилагательным и наречиям русского языка. Разрабатываемая База призвана отразить модели полисемии, характерные для русской признаковой лексики, т. е. предназначена для выявления возможных комбинаций значений в составе одной лексемы и типов семантических переходов между отдельными значениями.

В теоретическом плане задача семантического анализа русских прилагательных и наречий освещалась в докладе на конференции «Диалог'09» (Рахилина и др. 2009). В данной работе мы представим прикладной инструмент, позволяющий обобщить и систематизировать результаты теоретического исследования.

Изучение моделей полисемии и типов семантических сдвигов имеет богатые традиции как в отечественной, так и в зарубежной лингвистике. В то же время выявляемые в таких работах закономерности семантической деривации, как правило, основаны на рассмотрении отдельных сдвигов в лексемах самых разных классов и самых разных временных срезов. Задача «сплошного» обследования и системного анализа семантических переходов для большой и представительной группы лексики до сих пор не ставилась. При этом особенно мало внимания уделялось признаковым словам (исключениями в этом отношении являются работы Апресян 1974, Кустова 2004, Толстая 2008). Между тем исчерпывающий анализ и классификация каких-л. явлений становятся возможными только тогда, когда исследователь работает с достаточно полным перечнем таких явлений, иными словами, для изучения моделей семантической деривации необходимо создать представительный каталог таких моделей. Именно задачу инвентаризации моделей полисемии и типов семантических сдвигов на материале всей частотной адъективной и адвербиальной лексики и призвана решить создаваемая База данных.

2. Материал и методика исследования

Материалом для заполнения Базы послужили частотные многозначные качественные прилагательные русского языка (ср. ***веселая*** *девочка*), а также связанные с ними наречия (ср. ***весело*** *смеяться*) и предикативы (ср. *мне* ***весело***). Список таких

единиц был получен на основе Нового частотного словаря русской лексики (Ляшевская/Шаров 2008), составленного по данным Национального корпуса русского языка (НКРЯ). К частотным были отнесены прилагательные с встречаемостью не ниже 15 раз на миллион словоупотреблений. В соответствии с этим критерием было отобрано 300 полисемичных качественных прилагательных, которые вместе с соответствующими наречиями и предикативами и составили материал для Базы данных.

Семантический анализ вносимых в Базу единиц строился на основе изучения контекстов их употребления в НКРЯ. Для каждого прилагательного и наречия составлялся перечень таких контекстов, который затем подвергался классификации в соответствии с типом значения, реализуемом в данном контексте. Контексты, в которых признаковое слово выступает в одном и том же значении, обобщались: выделялся семантический параметр контекста, релевантный для выбора данного значения признакового слова. Так, группа контекстов слова *сладкий* в значении ‘обладающий сладким вкусом’, включающая, например, существительные *пирог*, *груша*, *блюдо*, *кисель*, *кофе*, объединяется семантическим признаком «еда и напитки», а контексты этого же слова со значением ‘доставляющий наслаждение’ — *печаль*, *тоска*, *чувство* — задаются параметром «эмоции» (ср. в этой связи практику разрешения семантической омонимии в НКРЯ, см. Рахилина и др. 2006, Шеманаева и др. 2007).

На следующем этапе анализа между всеми значениями признакового слова устанавливались связи и выявлялись семантические переходы, т. е. полисемия слова представлялась в виде семантической сети, в которой каждое последующее значение образовано от предыдущего посредством семантической деривации определенного типа. Общепринятыми в теоретических исследованиях являются два механизма сдвига значений: метафора и метонимия (ср. Lakoff/Johnson 1980, Croft 1993, Feyaerts 1999, Dirven 2002, Peirsman/Geeraerts 2006 и мн. др.). В нашем материале имеется множество примеров, демонстрирующих переходы обоих типов. Однако благодаря тому, что исследование строилось не на выборочном, а на «сплошном» анализе материала, в ходе работы наряду со стандартными случаями были выявлены переходы, не укладывающиеся в классические определения метафоры и метонимии, но вместе с тем обнаруживающие некоторые системные признаки. При семантических переходах такого рода имеет место **комбинация** метонимических и метафорических сдвигов параметров лексемы, основанная на конвенциализации импликатуры исходного значения (*дружно жить → дружно ненавидеть*; *крупно писать → крупно повезти*; *грубо вести себя → грубо ошибиться*). Переход такого рода получил название **ребрендинга** (подробнее см. Рахилина и др. 2009).

Важно заметить, что при построении семантических сетей отношения между значениями рассматриваются не отдельно для прилагательного и для связанного с ним наречия и предикатива, а в совокупности, как единая лексическая система, характеризующаяся общими принципами семантической эволюции. Часто именно такой подход позволяет обнаружить мотивированность деривационных моделей признаковой лексики (так, например, *теплый прием* естественно связать с *тепло принять*, а не с *теплый свитер*), ср. теорию метонимии, развиваемую в (Radden/Kövecses 1999), а также обсуждение соотношения между конструкциями вида наречие-глагол и прилагательное-отглагольное существительное в (Филипенко 2003: 101). Метонимические отношения между отдельными значениями разных частей речи (прилагательное-наречие или прилагательное-предикатив) классифицируются в нашей системе как транскатегориальная метонимия.

Построенная таким образом семантическая сеть переносится в Базу данных. К обсуждению ее структуры мы переходим в следующем разделе.

3. Структура Базы данных

База реализована с использованием SQLite (выбор СУБД обусловлен относительно небольшим объёмом данных и бесплатностью программного обеспечения). Для занесения значений в базу можно использовать любой из существующих редакторов баз SQLite, поддерживающий юникод, — например, SQLite2009 Pro. Поиск в базе осуществляется посредством обычных SQL-запросов, но для удобства пользователя нами была написана клиентская программа, позволяющая исследователям, не знакомым с языком SQL, задавать типичные запросы с помощью графического интерфейса, а также умеющая отображать результаты не только в виде таблицы, но и в виде части сети семантических переходов.

Каждая запись в Базе данных соответствует одному значению признакового слова. Формат Базы предполагает характеристику каждого отдельного значения по следующим параметрам:

1. **ID**: номер данного значения в Базе данных.
2. **lemma**: исходная форма описываемого слова (т. е. прилагательное в форме именительного падежа мужского рода единственного числа, наречие или предикатив).
3. **ID_parent**: номер значения, от которого образовано данное значение признакового слова.
4. **POS**: часть речи описываемого слова (прилагательное, наречие или предикатив).
5. **meaning**: краткое словарное толкование данного значения.
6. **short_description**: прототипический контекст для данного значения признакового слова, т. е.

своего рода «ярлык», наглядно представляющий характеризуемое значение. Так, для прилагательного *мягкий* в значении физического свойства объектов прототипический контекст задается существительным *диван*, а в значении качества человека — существительным *человек*. Этот «ярлык» используется, в частности, как идентификатор данного значения в семантической сети переходов, которую система может отобразить для каждого признакового слова. Тем самым пользователь, обращаясь к схеме значений, по конкретным примерам легко может определить, какие значения прилагательного или наречия скрываются за каждым узлом в сети.

7. **tax_class**: таксономический класс, к которому относится данное значение признакового слова. В основу классификации были положены признаки, используемые в семантической разметке НКРЯ. Но при этом по мере необходимости система пополняется новыми таксономическими категориями. Например, прилагательному *тяжелый* в сочетаниях типа *тяжелая сумка* приписывается характеристика ‘большой вес’; в сочетаниях типа *тяжелая арка* — ‘внешний вид’, *тяжелая голова* — ‘физическое состояние’, *тяжелая контузия* — ‘большая степень’ и т. п.
8. **ev**: оценочная характеристика данного значения признакового слова (положительная, отрицательная или нейтральная оценка). Этот параметр позволяет отслеживать, как развивается оценка в семантической парадигме лексемы.
9. **gram_restrict**: грамматические ограничения на функционирование слова в данном значении (например, для прилагательного *полный* в значении высокой степени, ср. *полное отчаяние*, недопустима краткая форма).
10. **context_sem**: семантические свойства контекста для данного значения признакового слова, т. е. семантические признаки существительных, которые могут сочетаться с прилагательным в данном значении, или семантические признаки глаголов, прилагательных и/или наречий, которые могут выступать при описываемом наречии в данном значении. Так, например, прилагательное *глубокий* в значении ‘имеющий большую глубину’ сочетается с конкретными существительными топологического класса «контейнеры» (*глубокая яма*), в значении ‘достигший предела в развитии’ — с существительными, принадлежащими к таксономическому классу «время» (*глубокая ночь*), а в значении ‘отличающийся глубиной содержания’ — с именами, относящимися к ментальной сфере (*глубокая мысль*). В некоторых значениях та или иная лексема может сочетаться одновременно с несколькими семантическими классами слов. Например, для прилагательного *легкий* в значении ‘невысокая степень’ определяемое существительное может принадлежать к одному из следующих семантических классов: «воздействие», «психическое или физиологическое состояние», «болезнь», «погодные условия», «температура», «свет», «цвет», «звук», «запах», «вкус» (ср. *легкое нажатие*, *легкий озноб, легкий ветер, легкий звон* и т. п.).
11. **sem_except**: список слов-исключений из правил, предписанных в context_sem. Этот параметр предназначен для таких случаев, когда прилагательное/наречие в целом может сочетаться с некоторым классом или классами существительных/глаголов, прилагательных или наречий (это правило задается в context_sem), но при этом среди единиц этого класса обнаруживаются такие, которые по каким-л. причинам нарушают общее сочетаемостное правило, т. е. не могут выступать с прилагательным/наречием в описываемом значении. Эти исключения и составляют значение данного параметра.

Так, с прилагательным *жестокий* в значении ‘являющийся проявлением сурового и беспощадного характера’ (метонимический переход от исходного значения — ‘свойство человека’) употребляются среди прочих существительные класса «речь», ср. *жестокие слова/ вопрос/ реплика*. В то же время при некоторых существительных данного класса слово *жестокий* получает другое значение — оно соотносится с идеей ‘высокой степени’, ср. *жестокий спор/ дискуссия/ перебранка*. По-видимому, релевантным в данном случае для противопоставления внутри таксономического класса «речь» оказывается признак направленности речевого действия. Семантика прилагательного *жестокий* в конструкции типа *жестокие слова* предполагает наличие субъекта (тот, кто проявляет жестокость) и адресата (тот, на кого направлена жестокость). Если существительное, выступающее при слове *жестокий*, обозначает речевое действие, производимое одним участником ситуации и обращенное к другому участнику (ср. *слова*, *вопрос*, *реплика*), то тем самым семантика такого существительного согласуется с «адресным» значением прилагательного. Если же существительное класса «речь» подразумевает реципрокальное отношение между участниками ситуации (ср. *дискуссия*, *спор*, *перебранка*), то происходит переинтерпретация значения прилагательного: оно указывает не на проявление жестокости участников ситуации по отношению друг к другу, а на интенсивность протекания ситуации в целом. Тем самым параметр sem_except, во-первых, позволяет уточнить характеристику семантического контекста для описания прилагательного в Базе и, во-вторых, имеет теоретическую значимость: он выявляет когнитивно значимые признаки лексических единиц языка.

12. **context_gram**: грамматические свойства контекста для данного значения прилагательного, наречия или предикатива. Например, одно из значений прилагательного *верный* — ‘надежный, преданный’ (ср. *верный друг*) — наряду со стандартной атрибутивной моделью может реализоваться в конструкции с дательным падежом, ср. *остаться верным кому-л*.
13. **trans_type**: тип семантического сдвига. В Базе предусмотрено 4 типа переходов:
    - метафора, ср. *сладкий пирог → сладкие мечты*;
    - метонимия, ср. *радостный мальчик → радостная музыка*;
    - транскатегориальная метонимия, т. е. метонимия, имеющая место между разными частями речи (см. раздел 2), в т. ч.
        - от некоторого значения прилагательного к некоторому значению наречия (*веселая девочка → весело смотреть*);
        - от некоторого значения наречия к некоторому значению прилагательного (*весело смотреть → веселый взгляд*);
        - от некоторого значения предикатива к некоторому значению прилагательного (*мне весело → веселая музыка*).
    - ребрендинг (см. раздел 2), ср. *страшный фильм* (= ‘фильм, внушающий страх’) → *страшная радость* (= ‘радость в большой степени’); *здоровый мальчик* (= ‘мальчик, наделенный здоровьем’) → *здоровая палка* (= ‘палка большого размера’); *смешной рассказ* (= ‘рассказ, вызывающий смех’) → *смешная порция* (= ‘маленькая порция’). Данные примеры невозможно охарактеризовать как метонимию или метафору в чистом виде. Во-первых, при этих переходах исходное и производное значение прилагательного принадлежат разным таксономическим классам, соответственно, их нельзя признать метонимическими (в современных теориях семантических сдвигов метонимия понимается как сдвиг в пределах одного таксономического класса, или домена, ср. Croft 1993, Radden/Kövecses 1999, Peirsman/Geeraerts 2006). Во-вторых, таксономические классы, соответствующие начальному и конечному значению, не обнаруживают концептуального сходства, которое могло бы служить основанием для метафорического сравнения, или проекции (mapping) элементов структуры исходной области на структуру производной (ср. Lakoff/Johnson 1980, Turner/Fauconnier 1995). Хотя идея сравнения и присутствует в семантических сдвигах такого рода, базовым механизмом для них является импликатура, при которой результат семантического перехода служит следствием, или выводом из исходного значения (подробнее о ребрендинге см. Рахилина и др. 2009, о некоторых результатах исследования ребрендинга при помощи Базы данных см. тж. ниже).
14. **example**: пример употребления прилагательного, наречия или предикатива в данном значении (из НКРЯ).

Таковы параметры, в соответствии с которыми характеризуется каждое значение признакового слова при внесении в Базу. Заполненная таким образом База данных открывает широкие возможности для изучения моделей многозначности признаковой лексики и типов семантической деривации.

4. Применение Базы данных: примеры запросов

Запросы к Базе могут строиться как по одному из вышеописанных параметров, так и по их различным комбинациям. Рассмотрим некоторые примеры.

К наиболее очевидным способам построения запроса является поиск по типу семантического перехода. Так, задав в качестве условия ‘тип перехода: метафора’, пользователь получит все содержащиеся в Базе примеры метафорических сдвигов в прилагательных и наречиях. Поиск по признаку ‘тип перехода: ребрендинг’ отобразит все семантические сдвиги, которые, по нашей классификации, не подпадают под канонические определения метафоры и метонимии. Анализ таких переходов показывает, что их результирующие значения соотносятся с ограниченным набором семантических классов, ср. ‘интенсивность’ (*большой: дом → плакса; страшный: рассказ → удивление*), ‘большое количество’ (*солидный*: *костюм → зарплата*; *сумасшедший*: *человек → деньги*), ‘малое количество’ (*скромный*: *жилище → доход*; *жалкий*: *зрелище → крохи*), ‘большой размер’ (*густой*: *суп → слой*; *здоровый*: *мальчик → палка*), ‘положительная оценка’ (*редкий*: *выстрелы → розы*; *знатный*: *гражданин → щи*), ‘отрицательная оценка’ (*неважный*: *вопрос → здоровье*; *жуткий*: *рассказ → почерк*), ‘одновременность’ (*дружный*: *семья → аплодисменты*) и нек. др. Очевидно, что многие значения, возникающие в результате ребрендинга, связаны между собой регулярными соотношениями, ср. интенсивность/большой размер/большое количество: одно и то же прилагательное может отсылать к разным значениям этого семантического комплекса — в зависимости от свойств определяемого им существительного.

Запрос к Базе может строиться с учетом исходного или конечного значения признакового слова. Так, например, можно узнать, какие значения в результате семантического сдвига могут получать прилагательные размера. База показывает, что они могут переходить в следующие таксономические классы:

- ‘длительность’ (*большой*: *дом → привал*; *короткий*: *веревка → путь*);

- ‘свойство звука’ (*тонкий*: *палка → голос*);
- ‘свойство множества’ (*большой*: *дом → семья; маленький*: *яблоко → отряд*; *мелкий*: *орех → группа*);
- ‘свойство абстрактного явления’ (*крупный*: *песок → успех*; *глубокий*: *колодец → знания*; *мелкий*: *камень → подробность*; *узкий*: *лента → тема*);
- ‘нефизическое свойство человека’ (*мелкий*: *камень → чиновник*; *большой*: *дом → начальник*).

И наоборот, можно выяснить, какие значения могут быть исходными для прилагательных, например, с семантикой ‘нефизическое свойство человека’: к ним относятся

- ‘физическое свойство объекта’ (*жесткий/мягкий*: *сидение → человек*);
- ‘физическое свойство человека’ (*сильный*: *человек → союзник*; *крепкий*: *мужик → хозяйственник*);
- ‘физическое свойство животного’ (*скользкий*: *угорь → человек*; *чуткий*: *зверь → человек*).

Ограничение можно строить и по контекстным условиям реализации значения признакового слова, т. е., например, для прилагательного — по таксономическому классу существительных, с которым оно может сочетаться в исходном или производном значении. При этом в запросе можно задать одновременно параметры контекста исходного и конечного значения и тем самым проверить, соотносятся ли в русском языке свойства двух выбранных концептуальных областей. Так, если в качестве контекста исходного значения задать признак ‘человек’, а производного — ‘организации’, то система выдаст следующие примеры переходов: *сильный / слабый*: *человек → государство*; *юный*: *девушка → город*, а запрос на сдвиги с контекстами ‘вещество’ → ‘эмоции’ вернет результаты *сладкий / горький*: *лекарство → мука*.

Предметом изучения на материале Базы могут стать и оценочные значения признаковой лексики. Хорошо известно, что признаковые слова часто получают оценочную семантику. Так, анализ примеров, выведенных Базой с помощью соответствующего запроса, показывает, что отсутствующая в исходном значении оценка может появляться в производном как результат метафорического переноса, ср. *громкий*: *голос → фразы*; *теплый*: *камень → слова*; *искусственный*: *жемчуг → смех*; *дешевый*: *мыло → популярность*. Однако в нашем материале нет метафорических переходов, характеризующихся исчезновением в производном значении отрицательной оценки, присутствующей в исходном, или ее смены на положительную. Напротив, при ребрендинге положительная и отрицательная оценка могут в значительной степени стираться, заменяясь на значение интенсивности или большого размера (*добрый*: *человек → кусок*; *страшный*: *фильм → радость*; *жуткий*: *рассказ → сладкоежка*; *сумасшедший*: *человек → деньги*; *дикий*: *крик → восторг*).

Посредством Базы можно проверять и уточнять некоторые теоретические положения современной теории метафоры и метонимии. Так, конституирующим свойством метафоры, по крайней мере начиная с классической работы Lakoff/Johnson 1980, считается мена таксономического класса лексемы, претерпевающей семантический сдвиг. Тем не менее в нашем материале обнаруживаются примеры метафор, при которых исходное и производное значение самого признакового слова соотносятся с одним и тем же таксономическим классом, метафорическое сравнение же возникает только за счет изменения класса определяемого объекта. Это явление характерно прежде всего для прилагательных, описывающих внешние признаки чего-л., т. е. метафора в данном случае возникает за счет внешнего сходства объектов разных концептуальных зон, ср. *горбатый*: *человек → мост*; *стройный*: *девушка → дерево*; *вялый*: *трава → кожа*; *рваный*: *одежда → рана*. Исключительный теоретический статус этих переходов вполне согласуется с их наивно-языковым восприятием. Интуитивно понятно, что «концептуальное расстояние» между исходным и производным значением прилагательного в данном случае меньше, чем при метафорических сдвигах, сопровождающихся меной таксономического класса признакового слова, ср. переход от физического к нефизическому свойству (*черствый*: *хлеб → человек*). Тем самым в некотором смысле можно говорить о метафоре как о градуируемом явлении, т. е. те или иные переходы можно оценивать как «более» или «менее метафорические».

Итак, разработанная База данных позволяет решать различные исследовательские задачи, нацеленные на семантическое описание русских прилагательных и наречий. В перспективе, как нам представляется, материал Базы мог бы стать и основой для лексико-типологических исследований в области признаковых слов.

Литература

1. *Апресян Ю. Д.* Лексическая семантика (синонимические средства языка). М.: Наука, 1974.
2. *Кустова Г. И.* Типы производных значений и механизмы языкового расширения. М.: Языки славянской культуры, 2004.
3. *Ляшевская О. Н., Шаров С. А.* Новый частотный словарь русской лексики. 2008. http://dict.ruslang.ru/freq.php
4. *Рахилина Е. В., Кобрицов Б. П., Кустова Г. И., Ляшевская О. Н., Шеманаева О. Ю*. Многозначность как прикладная проблема: Лексико-семантическая разметка в Национальном корпусе русского языка // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2006» (Бекасово, 31 мая — 4 июня 2006 г.) / Под ред. Н. И. Лауфер, А. С. Нариньяни, В. П. Селегея. — М.: Изд-во РГГУ, 2006.
5. *Рахилина Е. В., Карпова О. С., Резникова Т. И.* Модели семантической деривации многозначных качественных прилагательных: метафора, метонимия и их взаимодействие // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2009» (Бекасово, 27–31 мая 2009 г.). Вып. 8 (15). — М.: РГГУ, 2009.
6. *Толстая С. М.* Пространство слова. Лексическая семантика в общеславянской перспективе. — М.: Индрик, 2008.
7. *Филипенко М. В.* Семантика наречий и адвербиальных выражений — М.: Азбуковник, 2003.
8. *Шеманаева О. Ю., Кустова Г. И., Ляшевская О. Н., Рахилина Е. В.* Семантические фильтры для разрешения многозначности в Национальном корпусе русского языка: прилагательные // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2007» (Бекасово, 30 мая — 3 июня 2007 г.) / Под ред. Л.Л. Иомдина, Н.И. Лауфер, А. С. Нариньяни, В. П. Селегея. — М.: Изд-во РГГУ, 2007.
9. *Croft W.* The role of domains in the interpretation of metaphors and metonymy // Cognitive Linguistics, 4, 1993, pp. 335–370.
10. *Dirven R.* Metonymy and metaphor: Different mental strategies of conceptualisation //R. Dirven and R. Pörings (eds.) Metaphor and Metonymy in comparison and contrast. Berlin/New York: Mouton de Gruyter, 2002. Pp. 75–111.
11. *Feyaerts K.* Metonymic Hierarchies: The Conceptualization of Stupidity in German Idiomatic Expressions // K.-U. Panther and G. Radden (eds). Metonymy in Language and Thought. Amsterdam & Philadelphia: John Benjamins, 1999.
12. *Lakoff G., Johnson M.* Metaphors We Live By. Chicago: University of Chicago Press, 1980.
13. *Peirsman Y., Geeraerts D.* Metonymy as a Prototypical category // Cognitive Linguistics 2006, 17(3). Pp. 269–316.
14. *Radden G., Kövecses Z.* Towards a Theory of Metonymy // K.-U. Panther and G. Radden (eds). Metonymy in Language and Thought. Amsterdam & Philadelphia: John Benjamins, 1999. Pp. 17–59.
15. *Turner M.; Fauconnier G.* Conceptual Integration and Formal Expression. In: Metaphor and Symbolic Activity, 1995. 10:3. Pp. 183–203.

Диалектная лексикография: электронная картотека «Архангельского областного словаря»

Dialect lexicography: an electronic corpus of «The Arkhangelsk region dialect dictionary»

Качинская И. Б. (kacza@rambler.ru)
МГУ имени М. В. Ломоносова, Москва

Крылов С. А. (krylov-58@mail.ru)
Институт востоковедения РАН, Москва

«Архангельский областной словарь» (АОС) — крупнейший диалектный словарь одного региона, его словник насчитывает ок. 180 тыс. слов. Основой словаря является «бумажная» картотека — в ней более 5 млн карточек. В Электронной картотеке АОС содержится уже более 1 млн «карточек». Важной проблемой Электронной картотеки является совершенствование программ по лемматизации: по переходу от фонетической словоформы (первоначальные полевые записи осуществляются в фонетической транскрипции) к грамматической и далее к начальной форме слова. Доклад сопровождается презентацией с демонстрацией автоматизированной обработки материала: (а) перевода текста полевой тетради (rtf) в базу данных (dbf) с заполненными полями; (б) грамматическими словоформами, расставленными в алфавитном порядке; и (в) гипотетическими начальными формами, созданными автоматическим анализатором в среде StarLing.

1. «Архангельский областной словарь» (АОС) под редакцией О. Г. Гецовой — крупнейший диалектный словарь одного региона. Его архив содержит более 2 тыс. полевых тетрадей и ежегодно пополняется на 50–100 тетрадей, записанных студентами русского отделения филологического факультета МГУ в рамках полевой диалектологической практики, аспирантами и руководителями практики[1]. Количество «бумажных» карточек составляет ок. 5 млн (ежегодное пополнение — 20–40 тыс. карточек). Словник АОС включает ок. 180 тыс. слов[2]. В 12-м выпуске АОС закончился материал на букву Д[3]. Том на Е-Ж пришлось разделить на два выпуска, слова на букву **Ж** закончатся в 14-м вып. Предполагаемый общий объем издания — не менее 60 выпусков.

2. Традиционно обработка полевых тетрадей сводилась к созданию «бумажной» картотеки и включала следующие этапы: 1) расписывание полевых тетрадей (или расшифровок аудиозаписей) на карточки; 2) карточки расставлялись в алфавитном порядке; 3) выявлялись новые слова по Словнику АОС.

С 1996 г. началась работа по созданию Корпуса «Электронная картотека АОС»[4], в базе уже более 1 млн «карточек» (ок. 10 млн словоупотреблений). Создание Корпуса АОС позволило 1) обеспечить лучшую сохранность материалов АОС: бесценная, десятилетиями собираемая картотека — это бумажные карточки в деревянных каталожных ящиках; уже давно стояла проблема сохранности архива; 2) ча-

---

[1] Картотека «Архангельского областного словаря» и архив тетрадей хранятся в кабинете диалектологии кафедры русского языка филологического факультета МГУ.

[2] Обратный словарь архангельских говоров / Под ред. О. Г. Гецовой. М.: «Наука», 2006.

[3] Первые 9 выпусков АОС выставлены на сайте http://www.philol.msu.ru/~dialectology/dictionary/ в формате .pdf (филологический факультет МГУ, каф. русского языка, каб. диалектологии, Словарь).

[4] Работа была поддержана грантом РГНФ, проект № 02-04-12020 в «Электронная картотека “Архангельского областного словаря”». Участники проекта: И. Б. Качинская, С. А. Крылов, Ф. С. Крылов, С. Г. Болотов.

стично решить проблему постоянного дефицита места хранения новых карточек; 3) вносить добавления нового материала в готовящиеся к изданию выпуски АОС почти в «готовом» виде; 4) использовать материалы АОС в самых различных научных целях.

В память компьютера вводятся полевые тетради, в первую очередь нерасписанные, из экспедиций последних лет; постепенно вводятся расписанные тетради из архива; начинается ввод «старой» картотеки, уже использованной для создания первых выпусков АОС (в ней теперь видится много полезного материала, не учтенного в ранних выпусках); предполагается ввод нового материала (буквы А-Ж), поступившего в картотеку после выхода соответствующих томов[5].

Полевые тетради набираются студентами-русистами в рамках камеральной студенческой диалектологической практики. По каждой тетради создается своя база данных, после чего они объединяются в общую, сводную базу. Далее производится сортировка по словоформе размеченного диалектного слова или по ключевому диалектному слову, данному в орфографической записи, — хотя сортировка может осуществляться и по любому другому полю (например, по району записи, населенному пункту). На сегодняшний день обработано более 600 тетрадей, записанных в 72 нас. пунктах в большинстве районов Архангельской области. Основной приток нового материала для работы над АОС сейчас идет именно через Электронную базу данных (БД АОС).

Электронная картотека АОС создана на основе СУБД StarLing (автор — Сергей Анатольевич Старостин). Эта база делает возможной работу с фонетической транскрипцией любого уровня, т. к. позволяет включать произвольные шрифты, знаки и диакритики; позволяет сортировать материал в заданном алфавитном порядке: пользователь может произвольно включать любые знаки, заранее объявляя их последовательность или, напротив, приравнивая их друг к другу, например: е = ё или А = Á-ударное прописное = а = á-ударное строчное[6].

3. Предваряя онлайновую версию проекта «Вавилонская башня», С. А. Старостин писал: «Помимо этимологии и сравнительно-исторического языкознания, многолетний предмет моих штудий — автоматическая морфология русского языка. На этих страницах[7] вы имеете возможность ознакомиться с компьютерными базами данных по словарям Ожегова, Зализняка и Мюллера, а также проанализировать любое русское слово и получить его полную акцентуированную парадигму»[8].

Проблема лемматизации, во многом решенная для автоматической обработки текстов русского литературного языка, для диалектных текстов еще не решена. При традиционном подходе лемматизация обычно основывается на приведении заранее заданных словоформ *письменного языка* к начальной форме слова, заранее заданной в словарях литературного языка. Между тем основной текст в БД АОС — это текст в фонетической транскрипции, часто разного уровня, иногда с разными способами графической передачи близких фонетических явлений (*есть* = *éсьть* = *йéс′т′* = *йес′* = *jéс′*, для *есть-2* еще и *juc′* в разных графических вариантах — и т. д.). Кроме того, в распоряжении диалектологов регулярно оказываются записи, произведенные самими диалектоносителями, не всегда грамотными; эти записи также вводятся в память компьютера, необходима и их адекватная обработка; грамматические характеристики слова в говоре часто отличаются от таковых в литературном языке.

Поэтому следующим этапом работы с Электронной картотекой АОС стала работа по лемматизации — восстановлению начальной формы (леммы) в ее орфографическом варианте из фонетической словоформы. Это необходимо (а) для пополнения Словника АОС; (б) для передачи авторам нового материала из Корпуса АОС при написании словарных статей в следующие выпуски и для добавления нового материала в уже написанные словарные статьи; (в) для корректного поиска лексем при работе с Корпусом «Электронная картотека АОС». В 2003–2006 гг. авторам для работы над буквами Е, Ж было передано из электронной базы дополнение в 16,5 тыс. словоупотреблений. Почти все эти тысячи «карточек» пришлось обрабатывать вручную, заменяя непосредственно в Базе словоформу в фонетической транскрипции начальной формой в орфографической записи.

4. В первый период работы мы шли от словоформы (в фонетической транскрипции) — к начальной форме (в орфографической записи), т. е. от материалов полевых экспедиционных тетрадей, включенных в Электронный Корпус, к Словнику[9]. Для этого пришлось проделать значительную по объему предварительную работу. Было создано два «детранскриптора». Детранскриптор-1 переводит запись

---

[5] Задумано создание он-лайновой версии АОС с регулярным пополнением.

[6] СУБД StarLing, на основе которой создан Корпус «Электронная картотека “Архангельского областного словаря”», находится в открытом доступе в Интернете на сайте С. А. Старостина: http.starling.rinet.ru.

[7] Имеется в виду сайт проекта «Эволюция языка» — «Вавилонская башня».

[8] http://starling.rinet.ru/morpho.php?lan=ru

[9] Работа была поддержана грантом РГНФ, проект № 05-04-04274а «Создание грамматического словаря северных говоров (на базе транскрипционной записи устной диалектной речи)». Участники проекта: И. Б. Качинская, С. А. Крылов, Ф. С. Крылов.

в традиционной фонетической транскрипции, содержащуюся в полевых тетрадях, в «облегченную» фонетическую транскрипцию, принятую в изданиях АОС, и тем самым во многом унифицирует способ подачи записи. Детранскриптор-2 переводит фонетическую словоформу (в любом варианте фонетической транскрипции) в подобие грамматической орфографической формы (поле lex). В этом поле запускается грамматический анализатор Старостина–Зализняка[10]. Сначала анализатор опознал около 20 % словоформ, потом опознавал даже до 70 %, но в результате выяснилось, что большинство из них были опознаны неверно.

Пришлось учесть особую орфографию АОС, отличающуюся от стандартной орфографии. Это отсутствие *Ь* у существительных 3 скл. (*ноч, доч, мыш*); написание корней *раст-* как *рост-* (*ростъ*), *раб-* как *роб-* (*робутать*); приставок *раз-/рас-* как *роз-/рос-*; объединение приставок *при-* и *пре-* в одну приставку *при-* (*прикра́сной, прикрати́ться*); написание *е/ё* после мягких шипящих (*яищё, пальтецё*) и *о* после твердых шипящих (*жо́нка, жо́рнов*); написание *е* в суффиксе вместо *-иц-* в ЛЯ (*здоро́вьеце, отделе́ньеце*); наличие у субстантивов и слов со смешанным склонением окончаний *-ой/-ей* (вместо *-ый/-ий*) (*ле́шой, зимной, зимней, егорей, жабей, жо́лвей*) и проч. Программа учла эти отличия и приравняла их к формам в стандартной орфографии, имеющимся в Словаре Зализняка (*доч = дочь, василей = василий, яйцё = яйцо*).

5. Признав путь от словоформы к начальной форме слова на начальном этапе несколько преждевременным, мы пошли в обратную сторону: от начальной формы слова, зафиксированной в Словнике, к его грамматической словоформе, чтобы впоследствии уже новая программа «узнанную» грамматическую форму (а) возводила к начальной, зафиксированной в Словнике; (б) предлагала несколько начальных форм из Словника на выбор в случаях омонимии; (в) предлагала новую (гипотетическую) начальную форму, отсутствующую в Словнике[11].

На основе электронной версии Словника АОС был создан вариант словника с грамматической разметкой, т. е. с указанием частеречной принадлежности *всех* слов — в «Обратном словаре архангельских говоров» указана частеречная принадлежность всех служебных слов, местоимений и наречий; отметка о части речи у прилагательных, существительных и глаголов присутствует лишь в особых случаях. Снова был запущен грамматический анализатор Старостина-Зализняка, каждое слово получило свой словоизменительный индекс. Таким образом, была проведена индексация всех случаев, где Словник АОС совпадал с Грамматическим словарем А. А. Зализняка (а именно, повторены индексы у общерусских слов[12]). Велась активная работа по созданию программ, расширяющих возможности грамматического анализатора Старостина–Зализняка. Лексемы, не опознанные анализатором, размечались принудительно: были повторены индексы у всех префиксальных образований; назначены соответствующие индексы у собственно диалектных слов, имеющих определенные финали, с привлечением некоторых алгоритмов, которые удалось записать специальными строками в Программу. Работа велась главным образом по существительным (более 70 тыс. лексем), глаголам (более 63 тыс. лексем) и прилагательным (более 16 тыс. лексем).

Возникло много сложностей в связи с нечеткой разработанностью помет, сделанных для Обратного словаря. Так, например, в Словнике особо помечены все существительные на *-ой/-ей,* для того чтобы отличать их от прилагательных. Но при этом одинаково помеченными оказались существительные типа *зной, промой* (имеющие стандартное 2 скл. муж. рода) и *любе́зной, божа́той* (имеющие адъективное склонение). В случаях, когда слово имелось в Грамматическом словаре Зализняка, оно, конечно, опознавалось правильно (*зной*). Но таких случаев оказалось немного. Грамматическая информация, которая представлялась «лишней» для пользователя-лингвиста из-за своей «очевидности», для парсера такой очевидностью не обладает.

Словообразование в говорах происходит гораздо интенсивнее, чем в литературном языке, в т. ч. префиксальное и постфиксальное. Поэтому в программе произведены специальные записи, приравнивающие формы с различными префиксами к формам (с иными префиксами или без них), имеющимся в Словаре Зализняка (*добеспоко́иться = беспоко́иться, дозаставля́ть = заставля́ть*). Различие в постфиксах в случае их наличия (*гости́ться, дозаусыпля́ться*) оказалось не столь актуальным, т. к. модель Зализняка–Старостина практически везде порождает гипотетические постфиксальные формы, рассматривая их как пассив. В то же время каждый раз в качестве начальной формы она предлагает либо две глагольные формы (с *-ся* и без *-ся*), либо одну (без *-ся*). Этот момент оказалось легко преодолеть, прописав строку, сравнивающую словоформу с конечным *-ся* с начальной формой слова.

---

[10] Находятся в свободном доступе на http://starling.rinet.ru/downl.php?lan=ru#soft

[11] Работа была поддержана грантом РГНФ, проект № 08-04-12132в «Грамматический словарь северных говоров. Электронная версия (на базе Словника "Архангельского областного словаря")» Участники проекта: И. Б. Качинская, С. А. Крылов, Ф. С. Крылов.

[12] В Словник АОС, как и в любой диалектный словник или словарь дифференциального типа, включено большое количество общерусских слов, т. к. эти слова часто отличаются по своей семантике от аналогов в литературном языке.

При унификации индексов по окончаниям могли оказаться ошибочными указания на переходность или вид глагола (совершенный — несовершенный). Но, во-первых, вид и переходность для большей части словоизменительных таблиц не имеет значения; во-вторых, в большом количестве случаев вид диалектного глагола по словнику определить невозможно — необходимо обращаться к контексту.

От учета диалектной грамматики анализатором пока пришлось отказаться, т. к. во многих случаях это бы резко увеличивало омонимию и ничуть не уменьшало количество ручной обработки.

Например, в северных говорах у существительных встречается вариативность флексий *е/и* в ед. ч. Д.–П. I скл. (*к козы́, в Москвы́*), флексий *е/у/и* в ед. ч. П. п. II скл. (*в ле́се, в городу́, на кони́*), расширена сфера употребления флексии *-у* в ед. ч. Р.п. II скл. (*около го́роду, у мужику́)*, вариативны окончания мн. ч. в И., Р., Тв.; смешиваются парадигмы склонения «разносклоняемых» существительных (*день, путь, время, мать, дочь*), регулярна ориентация слов III скл. на I (*на пече́)* и т. д. Некоторые словоформы оказалось возможным задать анализатору «списком» (*бра́товья́, бра́ты́, бра́теи* в И. мн.; *бра́тьев, бра́тове́й, братовьёв, братовье́й, бра́те́й, бра́тье́й, бра́тее́й; сыновье́й, сыновьёв* в Р.-В. мн. и проч.). Списком даны словоформы личных и некоторых других местоимений.

Расширять Программы грамматического анализа, учитывающего омонимичные словоформы, на данном этапе мы отказались, т. к. основной нашей задачей пока является восстановление начальной формы слова по ее грамматическим вариантам. В этом случае варианты в системах, где Р. = Д. = П. (*из Москвы́, к Москвы́, в Москвы́*), дадут нам одну и ту же лемму (*Москва*). К случаям, когда происходит ориентация 3 скл. на 1-е (*на пече́*), когда для всех слов II скл. оказываются возможны Р-2 (*табака́* и *табаку́*) или П-2 (*в ле́се* и *в лесу́*), а также П-3 (*на кони́*), можно будет вернуться позже, учитывая конкретную частотность встретившихся в Базе словоформ. То же касается лексем с широким варьированием грамматических форм по говорам (слов типа *брат, сын*): готовые таблицы их словоизменения созданы и адаптированы уже не к Словнику, а к Электронному Корпусу.

Несмотря на высокую частотность и грамматическую подвижность некоторых лексем, подавляющее большинство слов, составляющих Словник АОС, реально зафиксировано в 1-3 контекстах. В дальнейшей работе над Грамматическим Словарем индексы будут постоянно уточняться — как в работе с контекстами из «бумажной» картотеки, так и из Электронной.

6. Для удобства работы была создана **программа перекодировки серии шрифтов** АОС, созданных специально для нужд Словаря еще в 90-е годы (Times New Roman AOS), в шрифты уникод — современные шрифты, позволяющие работать с диакритиками. Для этого в среде WW были созданы макросы: а) макрос по переброске шрифтов АОС в уникоды; б) макрос, необходимый для промежуточной обработки полевых экспедиционных тетрадей перед проверкой и помещением их в базу (где знак * заменяется на знак стандартного ударения, а также производятся еще некоторые замены).

Был создан **модуль для импорта полевых тетрадей в dbf**. Теперь процедура обработки полевой экспедиционной тетради в 90 страниц, подготовленной для передачи в Корпус, полностью автоматизирована и занимает ок. 1 минуты: цельный текст (файл rtf) разбивается на поля (в т. ч. поля, содержащие сведения: «фонетическая диалектная словоформа», «восстановленная начальная форма», «пример», «адрес записи», «год записи», «автор записи», «информант», «примечания» и нек. другие) и выстраивается по алфавиту словоформ в поле lex («начальная форма слова»).

7. В дальнейшем предполагается создать новый **Автоматический анализатор** — серию программ, которые позволят «отлавливать» нестандартные («окказиональные») грамматические (и фонетические) словоформы и новые слова в постоянно пополняющемся Корпусе АОС. Анализатор будет сравнивать эту словоформу с уже имеющимися — гипотетически построенными — словоформами, созданными на основе Словника АОС. По своим результатам программа должна напоминать систему для проверки правописания типа Spell Checker или «Орфо»: (а) она должна находить слова и словоформы, отсутствующие в Словнике АОС (и в гипотетических таблицах, построенных на основе Словника), (б) исходя из анализа «ошибки», предлагать пользователю возможность выбрать вместо фонетической словоформы орфографически «правильный» вариант написания (*поцанки* > *по́дсанки* или *паца́нки; ця́ица* > *ча́ица* или *ча́яться; ш* > предлог *с* или частицу *ж*); (в) отмечать и предъявлять новую словоформу: давать возможность пользователю пополнять словарь словоформ и слов, предлагая новую таблицу гипотетических словоформ и в дальнейшем учитывая эти новые сведения.

Референциальный выбор как многофакторный вероятностный процесс[1]

Referential choice as a multi-factor probabilistic process

Кибрик А. А. (aakibrik@gmail.com), Институт языкознания РАН

Добров Г. Б. (wslc@rambler.ru), МГУ им. М. В. Ломоносова

Залманов Д. А. (dm.zalmanov@gmail.com)

Линник А. С. (skylinnik@gmail.com), **Лукашевич Н. В.** (louk@mail.cir.ru), МГУ им. М. В. Ломоносова

Один из важнейших процессов, участвующих в порождении дискурса — референциальный выбор, то есть выбор языкового выражения при упоминании лица или объекта. Референциальный выбор зависит от большого числа одновременно действующих дискурсивных факторов. В докладе предлагается модель, основанная на методах машинного обучения и описывающая референциальный выбор в аннотированном корпусе английских текстов.

1. Вводные замечания

При порождении дискурса говорящие/пишущие постоянно сталкиваются с необходимостью упоминать те или иные лица или объекты, то есть осуществлять референцию. В естественном дискурсе примерно каждое третье слово так или иначе связано с референцией. Способность адекватно использовать референциальные выражения — одна из наиболее важных языковых способностей. В данной работе предпринимается попытка смоделировать референциальные процессы, происходящие при порождении дискурса.

2. Референциальный выбор

Среди типов референции наиболее частотным типом является конкретная определенная референция. В естественном дискурсе постоянно встречаются повторные и многократные упоминания конкретных определенных референтов. В таких случаях говорящий может использовать не только полные референциальные выражения (имена нарицательные, имена собственные, именные группы с модификаторми), но и редуцированные референциальные выражения, в первую очередь анафорические местоимения. Выбор референциального выражения говорящим для конкретного референта — это **референциальный выбор**.

Рассмотрим отрывок естественного письменного дискурса, иллюстрирующий референциальный выбор.

(86) Tandy said consumer electronics sales at its Radio Shack stores have been slow, partly because a lack of hot, new products. Radio Shack continues to be lackluster, said Dennis Telzrow, analyst with Eppler, Guerin Turner in Dallas. He said Tandy has done a decent job increasing sales by manufacturing computers for others and expanding sales of its Grid Systems Corp. subsidiary, which sells computers to bigger businesses, but it's not enough to offset the problems at Radio Shack. Sales at Radio Shack stores open more than a year grew only 2 % in the

[1] Данное исследование поддержано грантом № 09-06-00 390 Российского фонда фундаментальных исследований.

quarter from a year earlier, he said. As a result, Mr. Telzrow said he cut his fiscal 1990 per-share earnings estimate for Tandy to $ 4,05 from $ 4,2.

В этом отрывке статьи из Wall Street Journal повторно упоминаются три референта: компания Tandy Corp. (пять раз), магазины Radio Shack (четыре раза) и лицо по имени Dennis Telzrow (шесть раз). Из пяти упоминаний Tandy три осуществляются при помощи полных ИГ (конкретно, при помощи имени собственного), а два — при помощи местоимения *it*. Лицо Dennis Telzrow упоминается дважды посредством полных ИГ и четырежды посредством местоимения *he*.

От чего зависит референциальный выбор? Этому вопросу посвящена огромная литература, обозреть которую не представляется возможным; но см., например, Givón 1983, Fox 1987, Chafe 1994, Arnold 2008. Существует целый ряд работ, в которых авторы пытались выделить тот или иной единичный фактор как объясняющий референциальный выбор. Практика, однако, показала, что никакой единичный фактор не в состоянии описать все случаи референциального выбора. В работах Kibrik 1996, 1999 была предложена следующая модель. Референциальный выбор непосредственно зависит от статуса референта в когнитивной системе говорящего в данный момент. Когнитивный компонент, отвечающий за референциальный выбор — рабочая память. Если референт высоко активирован в рабочей памяти говорящего, то используется редуцированное референциальное средство. Напротив, если референт характеризуется низкой степенью активации, то используется полная ИГ. Имеются также промежуточные уровни активации, при которых возможно использование и полных, и редуцированных референциальных средств.

Уровень активации референта, в свою очередь, зависит от ряда характеристик референта и дискурсивного контекста. Эти характеристики именуются факторами активации. К числу важнейших факторов активации относятся: одушевленность референта; значимость референта в дискурсе (протагонизм); расстояние до предшествующего упоминания референта (антецедента), измеряемое в клаузах; наличие/отсутствие дискурсивной границы (например, границы абзаца в письменном тексте) между данной точкой и антецедентом; роль антецедента в своей клаузе (подлежащее, дополнение…) и некоторые другие.

Каждый фактор активации принимает в конкретном случае конкретное значение, которое вносит определенный вклад в интегральный коэффициент активации (КА). КА, таким образом, представляет собой равнодействующую всех факторов и непосредственно определяет референциальный выбор. Референциальный выбор зависит и от некоторых других компонентов, в частности от фильтра референциального конфликта (неоднозначности); подробнее см. указанные выше работы.

3. Арифметическая и нейросетевая модели референциального выбора

Общий подход, описанный в предыдущем разделе, был ранее реализован в виде двух математических моделей. Первая из этих моделей, которую можно назвать арифметической, была описана в работах Kibrik 1996, 1999 применительно к русскому и английскому нарративному дискурсу, соответственно. В этой модели КА варьировал примерно в диапазоне от 0 до 1, а числовые веса значений факторов были подобраны соответственно. Эти веса представляли собой десятичные дроби (от −0,4 до 0,7), а их взаимодействие было смоделировано как простое сложение. Диапазон КА был разделен на несколько интервалов. Так, согласно модели английской референции, при КА менее 0,3 непременно используется полная ИГ, при КА более 1,0 — непременно местоимение, и есть три промежуточных интервала, в которых и полные ИГ, и местоимения возможны, но обладают разной степенью предпочтительности.

В работах Grüning and Kibrik 2003, 2005 была предпринята попытка построить более математически адекватную модель, в которой вклад отдельных факторов в референциальный выбор определялся бы автоматически, а взаимодействие между факторами могло бы быть нелинейным. Эта модель была основана на методе нейросетей — одном из широко известных алгоритмов машинного обучения. Эта работа показала, что алгоритмы машинного обучения, подбирающие оптимальный набор параметров, влияющих на конечных выбор, в принципе пригодны для моделирования такого многофакторного процесса, как референциальный выбор.

Следует отметить, что в исследовании по нейросетевому моделированию референциального выбора общий когнитивный подход к референциальному выбору был существенно редуцирован. В отличие от арифметической модели, в которой имеется интегрирующий когнитивный компонент (коэффициент активации), в нейросетевой модели набор значений факторов непосредственно отображается на референциальный выбор. Это является недостатком данной модели, поэтому в работе Grüning and Kibrik 2005 была поставлена задача восстановить когнитивную адекватность в дальнейших исследованиях, основанных на методах машинного обучения.

Во всех упомянутых работах исследовались небольшие образцы дискурса, насчитывающие 100–150 референциальных выражений. Моделирование при помощи алгоритмов машинного обучения требует значительно больших объемов данных. В связи с этим была поставлена задача создания большого референциального корпуса.

4. Корпус для исследований по моделированию референциального выбора

Согласно результатам работ Kibrik 1996, 1999, в число факторов активации входит фактор **риторического расстояния** от текущей точки дискурса до антецедента. Понятие риторического расстояния (RhetD) основано на теории риторической структуры (Mann and Thompson 1988). Эта теория описывает иерархическую смысловую организацию дискурса. Каждая элементарная дискурсивная единица (чаще всего — клауза) является узлом риторической сети, терминальные узлы объединяются в группы по близости, а между узлами (как терминальными, так и внутренне сложными) устанавливаются отношения — симметричные (такие как последовательность или конъюнкция) или асимметричные (такие как причина, условие, уступка и т. д.). Как было показано в работе Fox 1987, близость к антецеденту по иерархической структуре дискурса не менее важна для референции, чем линейная близость. На основе этой идеи в Kibrik 1996 было предложено измерение RhetD, то есть расстояние от клаузы, в которой происходит референциальный выбор, до клаузы антецедента, подсчитанное в числе шагов по риторической сети. Эмпирические исследования показали, что этот фактор является наиболее мощным фактором референциального выбора. См. Kibrik and Krasavina 2005 о некоторых дискуссионных вопросах, связанных с подсчетом RhetD.

Фактор RhetD является одновременно наиболее «дорогим», поскольку для идентификации его значений для каждого референциального выражения необходима полная разметка риторической структуры дискурса — трудо- и времяемкое дело. В связи с этим в качестве корпуса для исследования референциального выбора крайне желательно было использовать корпус, в котором разметка по риторической структуре уже произведена. На время начала данного проекта (середина 2000-х гг.) такой корпус имелся ровно один: англоязычный корпус RST Discourse Treebank, созданный коллективом под руководством Д. Марку (http://www.isi.edu/~marcu/discourse/Corpora.html), см. Carlson et al. 2003.

Этот корпус включает 385 статей экономической или политической тематики из газеты Wall Street Journal, в этих статьях содержатся 176 383 словоупотреблений и 21 789 элементарных дискурсивных единиц. Пример фрагмента текста из корпуса был приведен выше (пример (1), текст №1374). На рис. 1 показан пример риторического графа, соответствующего фрагменту другого текста (№1315). В этом фрагменте можно видеть как симметричные отношения (Contrast, List), так и большое число асимметричных.

Тексты, входящие в RST Discourse Treebank, составили основу нового корпуса, в котором была осуществлена аннотация по целому ряду потенциальных факторов активации.

5. Референциальная аннотация

Референциальная разметка была произведена при помощи программы MMAX-2, специально созданной группой немецких компьютерных лингвистов для этих целей; см. http://mmax2.sourceforge.net/. Разметка в MMAX-2 осуществляется при помощи так называемой аннотационной схемы, включающей набор размечаемых параметров (факторов). Эта схема была составлена О. Н. Красавиной и К. Чиаркосом (Krasavina and Chiarcos 2007). Она основана на языке XML и устроена по принципу stand-off annotation (файлы с аннотацией отдельны от самих текстов). Схема включает три последовательных компонента:

- выбор аннотируемых элементов
- разметка анафорических связей между ними
- разметка дополнительных признаков аннотируемых элементов.

Аннотируемый, или маркируемый, элемент (*markable*), далее **маркабула**, **—** это составляющая текста, которая может быть референциальным выражением. В качестве маркабул выступают либо

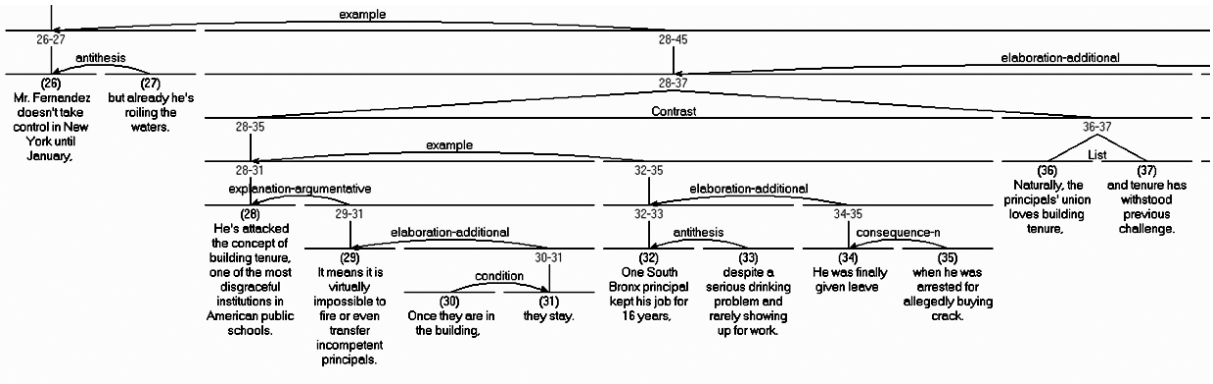


Рис. 1. Фрагмент риторического графа из корпуса RST Discourse Treebank

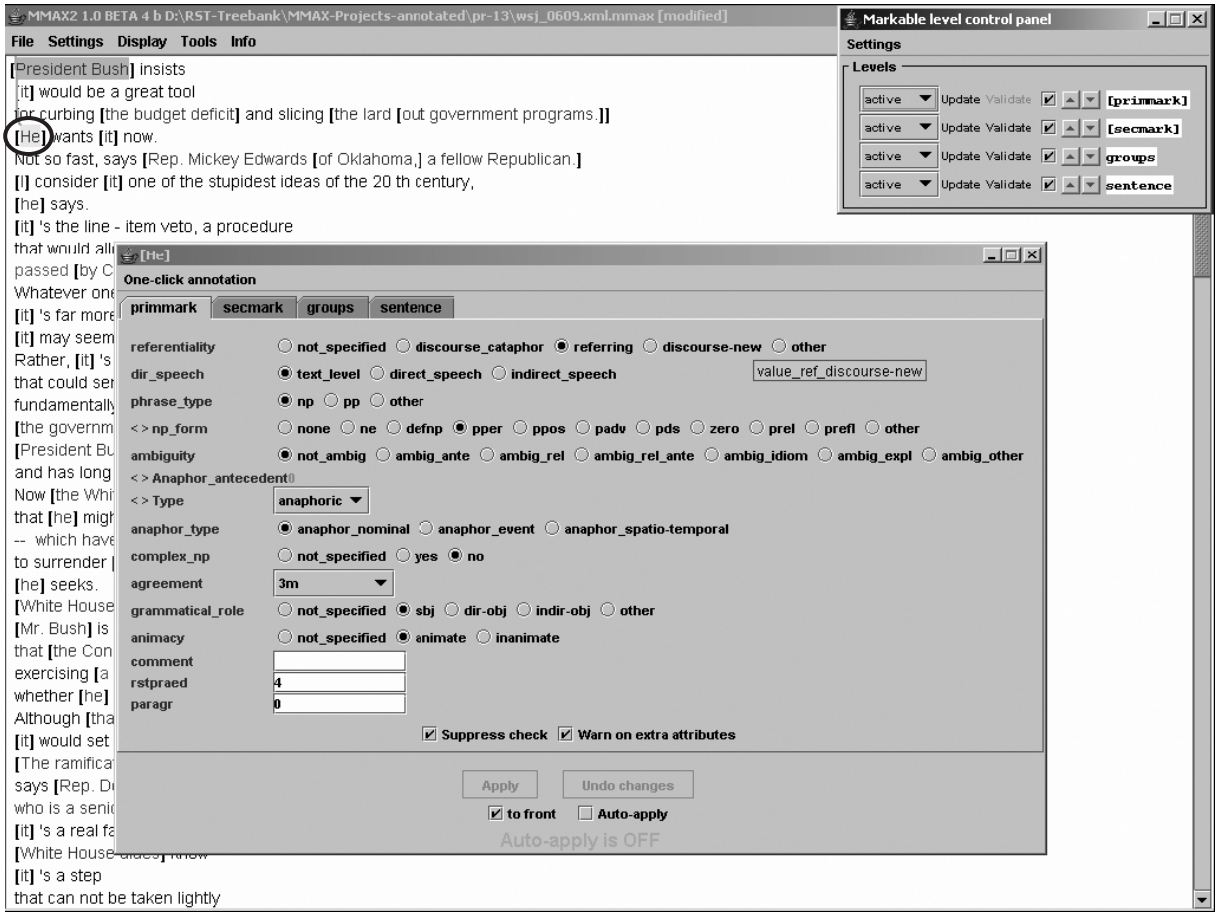


Рис. 2. Окна программы MMAX-2

именные, либо предложные группы. На рис. 2 можно видеть текстовое окно программы MMAX-2, в котором маркабулы обозначены при помощи квадратных скобок (текст №0609 корпуса). В рамках данной аннотационной схемы выделяется два типа маркабул:

- основные маркабулы, которые могут использоваться в анафорической функции; к ним принадлежат:
  - определенные, указательные и посессивные ИГ
  - имена собственные
  - личные и указательные местоимения
- второстепенные маркабулы, которые соответствуют дискурсивно-новым референтам и не могут употребляться в анафорической функции, однако могут являться антецедентами анафорических выражений; к ним относятся:
  - неопределенные ИГ (с неопределенным артиклем и без артикля)
  - элементы, которые не были классифицированы как основные маркабулы, но являются антецедентами основных маркабул.

Между выделенными маркабулами устанавливаются **отношения кореферентности**. Отношение кореферентности соединяет каждое непервое упоминание *n* некоторого референта с предшествующим ему упоминанием *n–1*, то есть антецедентом. Так, от местоимения [he] (выделено на рис. 2 при помощи овала) проходит линия к антецеденту — именной группе [President Bush].

Наконец, последний компонент схемы предполагает разметку **признаков** каждой маркабулы. Основные и второстепенные маркабулы имеют разные, но пересекающиеся наборы признаков. На рис. 2 в центре можно видеть рабочее окно программы MMAX-2, в котором показаны значения признаков для вышеупомянутой маркабулы [he]. Признаки соответствуют потенциальным факторам активации, таким как грамматическая роль или одушевленность; см. ниже.

Аннотация корпуса осуществлялась в основном вручную студентами-практикантами ОТиПЛ МГУ, однако перед этим тексты прошли автоматическую обработку. Так, автоматически выделялись маркабулы, которые легко обнаружить по формальным признакам: местоимения; практически все имена собственные; именные группы с артиклями. (Эта работа в значительной мере была проведена студенткой ОТиПЛ МГУ А. Антоновой при помощи частеречного парсера компании Cognitive Technologies; см. Антонова 2004.) На уровне анафорических связей автоматическая обработка была минимальной: связи устанавливались только для тех пар «анафор —

антецедент», где анафором являлось личное или притяжательное местоимение. Также автоматически присваивались значения некоторых признаков там, где это можно было сделать, опираясь на одно лишь лексическое наполнение маркабул. Частично автоматизированным был также процесс проверки конечных версий разметки.

В настоящее время аннотирование корпуса по референции осуществлено примерно на 64 %. Описанные ниже результаты основаны на 247 текстах, содержащих около 110 000 словоупотреблений. В этих текстах размечено 26 024 маркабул, в том числе 7097 имен собственных, 8560 определенных дескрипций, 1797 личных местоимений 3 лица. (К числу других маркабул относятся местоимения 1/2 лица, притяжательные местоимения, указательные местоимения, неопределенные ИГ и некоторые другие категории.) В данном сегменте корпуса на настоящий момент получено 3502 надежных пар «анафор — антецедент». Входящие в них анафоры разбиваются на классы так: имена собственные — 1541 (44 %), определенные дескрипции — 976 (28 %), местоимения — 985 (28 %).

Две версии корпуса — риторическая и референциальная — образуют в совокупности единый продукт, который мы называем «корпус RefRhet». Некоторые предварительные сведения о раннем варианте этого корпуса были представлены в работе Красавина 2006.

6. Факторы референциального выбора

Из референциальной и риторической разметок, а также из структуры текста как такового для каждого референциального выражения извлекаются значения следующих потенциальных факторов активации (в скобках указаны названия признаков, принятые в аннотационной схеме):

Признаки референта:
- первое/непервое упоминание в дискурсе (referentiality)
- одушевленность (animacy)
- протагонизм

Признаки антецедента:
- Входит ли в состав прямой речи (dir_speech)
- Тип синтаксической группы (phrase_type)
- Грамматическая роль (gramm_role)
- Референциальная форма (np_form, def_np_form)

Признаки анафора:
- Входит ли в состав прямой речи (dir_speech)
- Тип синтаксической группы (phrase_type)
- Грамматическая роль (gramm_role)

Расстояния между анафором и антецедентом:
- Расстояние в словах
- Расстояние в маркабулах между анафором и антецедентом
- Линейное расстояние в клаузах
- Риторическое расстояние в элементарных дискурсивных единицах

7. Методы машинного обучения при моделировании референциального выбора

Моделирование референциального выбора было осуществлено при помощи система Weka (http://www.cs.waikato.ac.nz/ml/weka/, см. Hall et al. 2009), в которой реализовано множество алгоритмов машинного обучения и автоматизирована оценка их качества. Для экспериментов были выбраны несколько алгоритмов машинного обучения, относящихся к разным типам: логические алгоритмы классификации и логистическая регрессия. При выборе алгоритмов деревьев решений и решающих правил (логических алгоритмов) мы руководствовались интерпретируемостью получаемых ими описаний классов в виде понятных человеку конструкций. В качестве таких классификаторов мы выбрали алгоритм деревьев решений C4.5 и алгоритм решающих правил JRip. Выбор логистической регрессии обусловлен двумя соображениями: во-первых, качество работы этого алгоритма несколько превосходит качество работы логических, а во-вторых, логистическая регрессия позволяет получить оценки вероятности принадлежности каждому из классов (см. ниже).

Для контроля качества использовалась процедура скользящего контроля:
1. Обучающее множество делится на 10 частей.
2. Затем классификатор строит решающую функцию по 9 частям из 10.
3. Построенная решающая функция тестируется на оставшейся десятой части.

Процедура повторяется для всех возможных разбиений, а результаты потом усредняются. Критерием выбора наилучшего набора признаков и алгоритма является аккуратность — отношение правильно предсказанных типов референциальных выражений к их общему количеству. Ср. недавнюю работу Greenbacker and McCoy 2009, в которой реализуется сходный подход.

8. Основные результаты

Пары анафор–антецедент, где анафор представлен полной ИГ (именем собственным или определенной дескрипцией), составляют 72 % в данной вы-

борке. Соответственно, классификатор будет иметь смысл, если будет показывать качество выше 72 %. При использовании всех вышеперечисленных признаков с помощью логистической регрессии удалось достичь уровня аккуратности предсказания 86,8 %. Логические алгоритмы показали качество чуть более 85 %. Приведем в качестве примера несколько правил, порожденных алгоритмом JRip:

- *(грамматическая роль антецедента= подлежащее) И (Риторическое расстояние ≤ 1,5) И (Расстояние в словах ≤ 7) => местоимение*
- *(Расстояние в словах ≤ 20) И (грамматическая роль антецедента = подлежащее) И (2-я модель протагонизма ≤ 0,117647) И (одушевленный) => местоимение*
- *(одушевленный) И (Расстояние в маркабулах ≥ 2) И (Расстояние в словах ≤ 11) => местоимение*

Кроме того, было осуществлено моделирование троичного референциального выбора: определенная дескрипция vs. имя собственное vs. местоимение. Этот выбор предсказать оказалось значительное труднее, показатели аккуратности снизились следующим образом: логистическая регрессия показала результат 76 %, логические алгоритмы — 74 %. Такое снижение уровня аккуратности не является удивительным, поскольку набор факторов активации изначально ориентирован на базовый референциальный выбор между полной и редуцированной ИГ. Для того, чтобы достаточно хорошо предсказывать различие между именем собственным и определенной дескрипцией, нужны дополнительные факторы. Следует отметить, что характер этих дополнительных факторов требует дальнейших исследований. В современных работах по референции вопрос о том, как говорящие выбирают между разновидностями полных ИГ, является малоизученным, см. Линник 2009.

9. Вероятностный характер референциального выбора

Как было отмечено выше, существует значительное количество референциальных выражений, которые выбираются не детерминированным, а вероятностным образом. В работе Kibrik 1999 данная проблема была подвергнута детальному анализу. При помощи многоэтапной экспериментальной процедуры, включающей опрос значительного числа носителей английского языка, каждому фактическому референциальному выражению было поставлено в соответствие «потенциальное референциальное выражение» — оценка того, какие референциальные средства в принципе могли бы быть использованы в данной точке дискурса. Имеется пять типов таких потенциальных референциальных выражений. В арифметической модели референциального выбора каждому из этих типов соответствует определенный интервал значений коэффициента активации.

В таблице 1 показаны 6 возможных соответствий между пятью потенциальным и двумя фактическими типами референциальных выражений, а также количественные данные по небольшому набору данных, на котором была основана работа Kibrik 1999.

Как экстраполировать эти результаты на корпус RefRhet? Учитывая, что статьи Wall Street Journal представляют собой хорошо отредактированные тексты, можно предположить, что категории (2) и (4) не вызывают проблем, и в таких случаях использованы оптимальные референциальные средства — полные ИГ и местоимения, соответственно. Однако остается категория (3) — случаи, в которых местоимение и полная ИГ могут быть использованы с равной вероятностью. В таких случаях предсказать фактический референциальный выбор практически невозможно. Даже идеальный алгоритм должен предсказывать референциальный выбор с некоторым количеством ошибок. Можно предположить, что целевой достижимый уровень аккуратности вряд ли превысит 90 %. Напомним, что на данный момент алгоритмы машинного обучения предсказывают референциальный выбор с уровнем аккуратности около 85 %, то есть отклоняются от фактического референциального выбора в 15 % случаев. Если вышеприведенные рассуждения верны, то достигнутый результат весьма близок к «потолку», в принципе возможному для такого вероятностного процесса, как референциальный выбор. Вопрос, однако, в том, совпадают ли те случаи, когда алгоритмы дают отличный от фактов прогноз, с теми

Таблица 1. Потенциальные и фактические референциальные выражения (по Kibrik 1999).

Потенциальные референциальные выражения	(1) Только полная ИГ (19 %)	(2) Полная ИГ, ?местоимение (21 %)	(3) Местоимение или полная ИГ (28 %)*	(4) Местоимение, ?полная ИГ (23 %)	(5) Только местоимение (9 %)
Фактические референциальные выражения	Полная ИГ (49 %)			Местоимение (51 %)	

* В том числе 9 % фактических полных ИГ и 19 % фактических местоимений.

случаями, когда предсказание действительно невозможно. Ответ на этот вопрос требует дальнейших исследований.

Один из алгоритмов машинного обучения, а именно логистическая регрессия, обладает свойствами, которые позволяют смоделировать вероятностный характер референциального выбора. Для каждого референциального выражения логистическая регрессия выдает количественное значение, которое можно считать оценкой вероятности использования местоимения в данной точке дискурса. Эта оценка является некоторым аналогом коэффициента активации — интегрального показателя, представляющего собой равнодействующую всех одновременно действующих факторов активации (см. раздел 3).

10. Заключительные замечания

В данной статье описана модель референциального выбора в английском тексте. Работа основана на корпусе RefRhet, включающем несколько сотен текстов, размеченных по риторической структуре и по целому ряду потенциальных факоторов референциального выбора (свойства референта, свойства антецедента, расстояние до антецедента и др.) Был использован ряд алгоритмов машинного обучения, способных предсказывать референциальный выбор (в первую очередь, выбор между полной и редуцированной ИГ) на основе имеющихся признаков. Алгоритмы продемонстрировали аккуратность предсказания референциального выбора в районе 85 %.

Таким образом, референциальный выбор зависит от множества одновременно действующих факторов и в этом смысле представляет собой **многофакторный** процесс. Кроме того, референциальный выбор является **вероятностным** процессом: существует достаточно большое число случаев, в которых говорящий (пишущий) может использовать более чем одну референциальную опцию. Учитывая это обстоятельство, очевидно, что референциальный выбор не может быть предсказан с аккуратностью 100 %. Полученные на данный момент результаты предсказания довольно близки к теоретически возможному максимуму аккуратности.

Описанный в данной работе теоретический и методологический подход может быть применен к широкому кругу других языковых процессов, которые являются многофакторными и вероятностными — например, к выбору порядка слов, к выбору грамматического оформления предиката, к выбору просодических характеристик и т. д.

Литература

1. *Антонова А.* 2004. Алгоритм определения антецедентов анафорических местоимений и автоматическая референциальная разметка корпуса газетных статей Wall Street Journal. Курсовая работа. МГУ им. М. В. Ломоносова.
2. *Красавина О. Н.* 2006. Корпусно-ориентированное исследование референции (принципы аннотации и анализ данных). Дисс. ... кандидата филологических наук. М.: МГУ им. М. В. Ломоносова.
3. *Линник А. С.* 2009. Выбор именной группы в качестве референциального средства в зависимости от риторического расстояния до антецедента и уровня активации референта. Курсовая работа. МГУ им. М. В. Ломоносова.
4. *Arnold, Jennifer.* 2008. Reference Production: Production-internal and addressee-oriented processes // Language and Cognitive Processes 23,4: 495–527.
5. *Carlson, Lynn D., Daniel Marcu and Mary Ellen Okurowski.* 2003. Building a discourse-tagged corpus in the framework of Rhetorical Structure Theory // Jan van Kuppevelt and Ronnie Smith (eds.) Current directions in discourse and dialogue. Dordrecht: Kluwer, 85–112.
6. *Chafe, W. L.* 1994. Discourse, consciousness, and time. Chicago: University of Chicago Press.
7. *Fox, B.* 1987. Discourse structure and anaphora in written and conversational English. Cambridge: Cambridge University Press.
8. *Givón, T.* 1983. Topic continuity in discourse: An introduction // T. Givón (Ed.), Topic continuity in discourse: A quantitative cross-language study. Amsterdam: Benjamins, 1–42.
9. *Greenbacker, Charles F., and Kathleen F. McCoy.* 2009. Feature selection for reference generation as informed by psycholinguistic research // Production of referring expressions (PRE-CogSci) 2009: Bridging the gap between computational and empirical approaches to reference. Proceedings of the conference, Amsterdam, July 29, 2009.
10. *Grüning, André, and Andrej A. Kibrik.* 2003. A neural network approach to referential choice // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции Диалог-2003. М.: Наука, 260–266.

11. *Grüning, André, and Andrej A. Kibrik.* 2005. Modeling referential choice in discourse: A cognitive calculative approach and a Neural Networks approach // António Branco, Tony McEnery and Ruslan Mitkov (eds.). Anaphora processing: Linguistic, cognitive and computational modelling. Amsterdam: Benjamins, 163–198.
12. *Hall, M., F. Eibe, G. Holmes, B. Pfahringer, P. Reutemann, Ian H. Witten*. Witten. 2009. The WEKA data mining software: An update. SIGKDD Explorations, Volume 11, Issue 1.
13. *Kibrik, A. A.* 1996. Anaphora in Russian narrative discourse: A cognitive calculative account // B. Fox (ed.) Studies in anaphora. Amsterdam: Benjamins, 255–304.
14. *Kibrik, A. A.* 1999. Reference and working memory: Cognitive inferences from discourse observation // Discourse studies in cognitive linguistics. Ed. by K. van Hoek, A. A. Kibrik and L. Noordman. Amsterdam: Benjamins, 29–52.
15. *Kibrik, A. A., and Olga N. Krasavina.* 2005. A corpus study of referential choice: The role of rhetorical structure // Диалог. Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции Диалог'2005. Ред. И. М. Кобозева, А. С. Нариньяни, В. П. Селегей. М.: Наука, 561–569.
16. *Krasavina, Olga, and Christian Chiarcos,* Ch. 2007. PoCoS — Potsdam Coreference Scheme // Proceedings of the Linguistic Annotation Workshop (LAW). June 28–29, 2007, Prague, Czech Republic. Stroudsburg, PA: Association for Computational Linguistics, 156–163.
17. *Mann, William C., and Sandra A. Thompson.* 1988. Rhetorical Structure Theory: Toward a functional theory of text organisation // Text 8(3): 243–281.

Автоматическое формирование базы сочетаемости слов на основе очень большого корпуса текстов

Automatic construction of word combination database using huge text corpora

Клышинский Э. С. (klyshinsky@mail.ru),
Кочеткова Н. А. (natalia_k_11@mail.ru),
Литвинов М. И. (promithias@yandex.ru)
Московский государственный институт электроники и математики

Максимов В. Ю. (vadimmax2000@mail.ru)
Институт прикладной математики им. М. В. Келдыша РАН

В статье рассмотрены вопросы автоматического формирования базы сочетаемости слов (глагол или деепричастие + существительное, прилагательное + существительное, причастие + существительное) на основе анализа размеченного корпуса большого размера — более 109 словоупотреблений. Данная работа выполнена при финансовой поддержке ФЦП «Научные и научно-педагогические кадры инновационной России» на 2009–2013 годы.

Введение

Информация о взаимном сочетании слов является достаточно важной для задач анализа текстов на естественном языке. Обладая подобной информацией, можно, например, существенно повысить качество и скорость синтаксического анализа, причем как глубинного, так и поверхностного. Подобная база может использоваться для определения меры близости текстов, их содержания, при снятии омонимии и так далее.

Наиболее часто для этих целей используются коллокации, то есть устойчивые словосочетания. Устойчивость подобных сочетаний во многом является трудно определяемой величиной. В различных задачах под ними могут пониматься фразеологизмы, идиомы и несвободные словосочетания [1]. В связи с этим для практических задач используются статистические методы определения связности соседних слов, например, MI [2], t-score [3] и некоторые другие. Полученные таким образом коллокации могут использоваться для составления информационных портретов [4], снятия омонимии [5], выделения понятий предметной области [6], кластеризации документов, генерации связанных текстов и целом ряде других задачах.

Однако важную роль при анализе текстов играют и свободные словосочетания. Подобная информация также может помочь при решении многих задач. В связи с этим отечественные лингвисты уже длительное время ведут работы по созданию подобных словарей. На данный момент разработаны весьма представительные словари, как в бумажном [7], так и в электронном виде [8]. Однако объем подобных словарей с точки зрения машинной обработки текста прискорбно мал. Так, например, [7] содержит в себе всего 2500 статей, хотя и весьма представительных, приводящих не только информацию о сочетании слова с другими, но и толкования данного слова, его грамматические характеристики. В работе [8] приводится более 10 000 статей, что охватывает порядка 3–5 % современной русской морфологии. В области составления словарей коллокаций также объемы полученных результатов не слишком велики. Так, например, в работе [5] указывается, что был составлен словарь на 30 000 коллокаций. Зачастую это связано с тем, что составляются словари конкретных предметных областей и задач, а работы для языка в целом практически не проводятся.

Временные затраты на создание подобных словарей достаточно велики, что собственно и объясняет небольшой объем. В связи с этим встает во-

прос автоматизации процесса создания подобных словарей. Кроме того, для их применения в машинной лингвистике необходимо приведение словарей к представлению, удобному для машинной обработки. При этом вопрос стоит о привлечении и изучении больших объемов текста.

Ранее уже предпринимались попытки извлечения информации о глагольном управлении и глагольном примыкании из больших корпусов, а также для других видов сочетаемости слов. Однако работа с большим корпусом путем его просмотра требует огромного количества времени. Так, например, работы, проводимые Большаковым И. А. в течение 20 лет, позволили ему получить базу сочетаемости для 185 тысяч слов и выражений, в том числе около 57 тыс. титулов словаря для существительных (раздельно для единственного и множественного числе) и 38 тыс. титулов для глаголов (раздельно для инфинитива и личных форм). Общее количество сочетаний превышает 1,75 млн [9].

Объем и сложность проведенных работ объясняется среди прочего наличием различного рода неоднозначностей. Их наличие требует проведения ручной или автоматизированной (но ни в коем случае не автоматической) разметки, по результатам которой из текста выделяются, например, глаголы и зависимые от них слова. Дополнительные трудности вводит синтаксическая неоднозначность, в связи с чем для каждого предложения необходимо предварительно построить дерево зависимостей (хотя бы и мысленно), и на его основании пополнить базу глагольного управления.

Метод выделения сочетаемости слов

Для практических задач зачастую хватает информации о том, что данный глагол может употребляться с данным существительным. Для автоматического определения таких связей необходимо решить проблему лексической и синтаксической неоднозначностей. Для этих целей были выдвинуты две гипотезы. Первая из них (как нам представляется — наиболее сильная) состоит в том, что в тексте достаточно большого объема группы из однозначных с точки зрения морфологического анализа слов будут встречаться достаточно часто, чтобы собрать статистически значимые результаты. Под однозначностью здесь мы понимаем случай, когда в результате морфологического анализа слова возвращается единственная строка его нормальной формы. В связи с тем, что в русском языке для большинства слов имеется достаточное количество форм, вероятность обнаружить однозначное слово относительно велика. A priori основной вопрос заключался в том, насколько часто в тексте будут встречаться группы подобных слов. Вторая гипотеза состояла в том, что некоторые группы слов могут быть синтаксически однозначно подчинены другим словам даже без проведения синтаксического анализа. В соответствии со сформулированными гипотезами для генерации базы глагольной сочетаемости нами были использованы следующие простые положения.

1. Следующая за единственным глаголом группа существительного синтаксически подчиняется данному глаголу.
2. Единственная группа существительного, расположенная в начале предложения перед единственным глаголом, синтаксически подчиняется данному глаголу.
3. Прилагательные, расположенные перед первым в предложении существительным или между глаголом и существительным, синтаксически подчиняются данному существительному.
4. Положения 1–3 могут быть применены к деепричастиям и причастиям.
5. В тексте на русском языке должно быть представлено достаточно большое количество неомонимичных групп.

Само расположение выделенных групп с большой (но не стопроцентной) вероятностью позволяет говорить о корректности определения синтаксических зависимостей. Отсутствие неоднозначности гарантирует корректность определения нормальной формы слов. И, наконец, корпус текстов большого объема может гарантировать статистическую значимость результатов.

Итак, для рассмотрения были отобраны синтаксические конструкции, включающие глагол и единственную группу существительного перед ним или первую группу существительного после него. При этом группа существительного описывалась следующим образом: предлог притяжательное_местоимение числительное группа_прилагательных существительное. Все части группы существительного являются необязательными, а притяжательные местоимения и числительные игнорируются. Также отбрасывались и наречия. Точность результатов при этом будет определяться точностью выбора неомонимичных слов, корректностью выбора последовательности слов и вероятностью правильного применения второй гипотезы. Представительность результатов определяется объемом анализируемого корпуса и вероятностью встретить неомонимичную группу с заданными характеристиками.

Итак, для создания базы сочетаемости слов необходимо проанализировать корпус текстов большого размера, выделяя из него последовательности слов, отвечающие предложенным шаблонам. Для каждой уникальной последовательности должна быть подсчитана ее встречаемость, которая в дальнейшем используется для определения статистической значимости результата.

Описание эксперимента

Для экспериментов использовался полученный ранее и несколько обновленный и расширенный корпус текстов [10]. В качестве основы корпуса были использована Библиотека Мошкова, включающая в себя порядка 680 млн словоупотреблений. Кроме того, была использована еще одна коллекция художественной прозы, объемом около 120 млн словоупотреблений, включающая в себя как классических, так и современных авторов. Также использовалась новостная лента, опубликованная на сайтах РБК, Лента.ру, Российская и Независимая газеты, РИА Новости (всего более 325 млн. словоупотреблений), и новостные ленты околокомпьютерной тематики Компьюлента.ру и PCWeek (37 млн словоупотреблений). Конкретный объем каждого из источников приведен ниже в таблице. Общий объем корпусов составил почти 1,2 млрд словоупотреблений.

Источник	Объем, млн словоупотреблений
Библиотека Мошкова	680
РИА Новости	156
Доп. корпус прозы	120
Независимая газета	89
Лента.ру	33
Российская газета	29
PCWeek	28
РБК	21
Компьюлента	9
Итого	**1165**

Все полученные комбинации слов сохранялись в базе данных, работа с которой заняла основное время эксперимента. Для морфологического анализа использовался модуль морфологического анализа «Кросслятор» [11].

По результатам экспериментов были получены базы сочетаемости глаголов и существительных, деепричастий и существительных, существительных и прилагательных, существительных и причастий. Для каждого указанного типа сочетаний подсчитывалось общее количество их вхождений, то есть сколько раз в корпусе встретились сочетания данного типа. Кроме того, подсчитывалось количество уникальных сочетаний данного типа. Объем получившихся баз приведен ниже в таблице. Числитель показывает общее количество обнаруженных вхождений, знаменатель — количество уникальных сочетаний. Дополнительно подобный подсчет был осуществлен для сочетаний, встретившихся в корпусе более одного и двух раз (третий и четвертый столбец, соответственно).

Исследование результатов показало, что в выделенных парах приняло участие 21 500 глаголов из 26 400, представленных в морфологическом словаре, 53 300 существительных из 83 000, представленных в морфологическом словаре и 23 700 прилагательных из 45 300 имеющихся. Большое количество глаголов объясняется гораздо меньшей степенью их омонимичности. Низкое количество прилагательных объясняется тем, что из нескольких прилагательных, стоящих перед существительным, в базу помещалось только первое.

Пара	Всего вхождений, млн	> 1 повторения, млн	> 2 повторений, млн
Глагол+сущ.	65 / 8,3	60,3 / 3,5	57,7 / 2,3
Деепр.+сущ.	3,5 / 0,88	2,8 / 0,31	2,6 / 0,18
Сущ.+прил.	9,9 / 1,3	9,2 / 0,56	8,8 / 0,36

Наибольшая повторяемость сочетаний была достигнута на новостных текстах. Наиболее часто встречающими сочетания глагола и существительного оказались следующие (слова приведены к нормальной форме, рассматривается встречаемость во всех формах).

Сочетание	Встречаемость
Сообщить РИА	624 691
Передавать корреспондент	327 903
Покачать голова	304 597
Принять участие	271 250
Иметь в вид	201 167
Принять решение	140 090
Говориться в сообщении	132 385
Сообщать агенство	118 615
Сказать собеседник	115 959
Идти речь	108 306

Наиболее часто встречающимися глаголами в различных сочетаниях стали следующие (результаты округлены до тысяч).

Сочетание	Встречаемость
Быть	2 590 000
Сказать	1 908 000
Сообщить	1452 000
Иметь	721 000
Применять	623 000
Получить	525 000
Заявить	515 000
Передавать	462 000
Идти	450 000
Сообщать	425 000

Среди сочетаний существительное + прилагательное наиболее часто встречающимися оказались следующие.

Сочетание	Встречаемость
Ближайший время	23 664
Правый рука	19 809

Сочетание	Встречаемость
Официальный представитель	19489
Последний время	18555
Военный служба	17933
Большой количество	17737
Официальный сайт	17385
Левый рука	16121
Информационный агенство	15503
Молодой человек	14699

Наиболее часто встречающимися существительными в парах прилагательное + существительное в различных сочетаниях стали следующие.

Сочетание	Встречаемость
Время	69438
Голос	61409
Человек	58313
Рука	54467
Жизнь	52125
Система	49855
Количество	48821
Свет	46438
Место	42973
Работа	41741

Полученные результаты до некоторой степени соотносятся с имеющимися данными о частотах распределения слов русского языка. В значительной мере здесь чувствуется влияние новостной лексики.

Анализ показал, что в результаты не попали принципиально неоднозначные слова, такие, например, как «красный», выступающий как в роли прилагательного, так и в роли существительного. Кроме того, в базу не вошли устаревшие и чрезвычайно редко употребимые слова, например, «взгревать», «издаиваться», «парагвайка» и так далее.

Следует заметить, что в связи с неоднозначностью в рамках данного эксперимента в базу не попали целые пласты сочетаний. Так, например, омонимичными является большое количество предлогов, например, «при» (повелительное наклонение единственного числа от «переть»), «для» (деепричастие от «длить») и так далее. Однако подобная ситуация может быть исправлена достаточно легко введением фильтров. Автоматически брать все слова, которые могут быть предлогами, как предлоги было бы не всегда корректно. Так, например, слово «сверху» может выступать как в роли предлога, так и наречия, причем примерно равновероятно. С другой стороны, языковые конструкции, в которых данное слово встречается, существенно отличаются в зависимости от того, применяется в них наречие или предлог. В связи с этим в конечный результат будут включаться только «правильные» конструкции. Резюмируя можно сказать, что данный вопрос нуждается в дальнейшем исследовании.

Выборочный просмотр результатов показал, что количество ошибок не превышает 1 %. В области наиболее частотных сочетаний ошибки метода составляют порядка 0,1 %, тогда как сочетания, встретившиеся только один раз, выделяются с примерно 1–2 % ошибок. Часть из ошибок объясняется не совсем корректной обработкой некоторых видов конструкций. Так, например, в предложении «Хочу от лица коллектива поздравить юбиляра» конструкция «от лица» ошибочно относилась к глаголу «хотеть». Отдельную проблему представляют ассоциации, гиперболы и другие выразительные средства литературного языка. Так, например, конструкция «секретарша ускакала» хотя и выделяется правильно, но с трудом может быть названа характерной для поведения секретарей. С другой стороны, подобная конструкция встретилась в корпусе всего один раз и с очевидностью находится ниже уровня статистической значимости. Будучи оторванными от контекста, подобные конструкции удивляют, хотя их выделение с точки зрения приведенных выше шаблонов проводится вполне корректно. С другой стороны, в текстах одной из новостных лент регулярно встречались выражения вида «сообщает в четверг». Несмотря на серьезные возражения о стилистической корректности подобного сочетания, его выделение проводилось вполне корректно. Будучи же приведенным к нормальной форме («сообщать в четверг»), выражение не вызывает никаких возражений.

Наиболее частотные сочетания, полученные подобным образом, хорошо коррелируют с результатами, получаемыми методами выделения коллокаций (здесь автор хотел бы выразить признательность Ягуновой Е. В. и Пивоваровой Л. М. за проведенное сравнение). Так, например, в обоих методах самым встречающимся сочетанием в новостных текстах было «агентство сообщать».

Выводы

Приведенный в работе метод позволяет на больших объемах текстов получить приемлемые результаты по извлечению глагольной сочетаемости. Несмотря на то, что для построения баз было использовано около 1,5 % всех словоупотреблений, большой объем корпуса позволил получить представительный результат.

Проведенные эксперименты показали, что выдвинутые гипотезы вполне корректны, хотя и носят вероятностный характер. При этом точность получаемых результатов составляет порядка 99 %. Отдельной темой для исследований является корректность встречающихся конструкций с точки зрения правил, принятых в языке.

Полученный корпус глагольной сочетаемости позволит перейти к следующим экспериментам в области сочетаемости слов: снятие неоднозначностей, группировка слов по семантическим признакам и так далее. Планируется провести аналогичные эксперименты на корпусе со снятой статистическими методами омонимией. Это позволит значительно увеличить базу обрабатываемых сочетаний и, как следствие, существенно увеличить объемы базы. С другой стороны, это должно привести к увеличению процента ошибок.

Литература

1. *Хохлова М. В.* Экспериментальная проверка методов выделения коллокаций // Сб. статей «Инструментарий русистики: корпусные подходы». — Хельсинки, 2008. С. 343–357
2. *Church K., Hanks, P.* Word association norms, mutual information, and lexicography, *Computational Linguistics*, 1990, 16(1), P. 22–29.
3. *Stubbs, M.* Collocations and semantic profiles: On the cause of the trouble with quantitative studies, 1995. *Functions of Language*, 1.
4. *Антонов А. В., Ягунова Е. В.* Лингвистический анализ информационного портрета как свертки множества текстов. Постановка эксперимента // Сб. трудов научно-практического семинара «Новые информационные технологии в автоматизированных системах-13». М.: МИЭМ, 2010. С. 50–59.
5. *Невзорова О. А., Невзоров В. Н., Зинькина Ю. В., Пяткин Н. В.* Интегральная технология разрешения омонимии в системе анализа текстовых документов «ЛоТА» // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2007» М.: Изд-во РГГУ, 2007, С. 422–427
6. *Большакова Е. И., Баева Н. В., Бордаченкова Е. А., Васильева Н. Э. Морозов С. С.* Лексико-синтаксические шаблоны в задачах автоматической обработки текста // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2007» М.: Изд-во РГГУ, 2007, С. 70–75
7. *Словарь* сочетаемости слов русского языка / Под ред. П. Н. Денисова, В. В. Морковкина. 3-е изд., испр. М., АСТ, 2002. 816 с.
8. *Бирюк О. Л., Гусев В. Ю., Калинина Е. Ю.* Словарь глагольной сочетаемости непредметных имен русского языка — http://dict.ruslang.ru/abstr_noun.php
9. *Большаков И. А.* Кросслексика — большой электронный словарь сочетаний и смысловых связей русских слов // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2009». Вып. 8 (15). — М.: РГГУ, 2009. 620 с.
10. *Клышинский Э. С.* Некоторые сложности автоматизированной лемматизации несловарных словоформ // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2009». Вып. 8 (15).– М.: РГГУ, 2009, С. 165–169.
11. *Елкин С. В., Клышинский Э. С., Стекляников С. Е.* Проблемы создания универсального морфосемантического словаря // Сб. трудов Международных конференций IEEE AIS'03 и CAD-2003. Т. 1. Дивноморское. 2003

Поиск хозяина предложной группы в русском предложении[1]

Search of the control word of a prepositional phrase in russian

Кобзарева Т. Ю. (stamstam@mtu-net.ru)

Российский государственный гуманитарный университет

Рассмотрены проблемы поиска хозяина предложной группы в русских текстах разных жанров при автоматическом поверхностно-синтаксическом анализе. Описывается лингвистический базис и стратегия поиска хозяина предлога в рамках системы модульного анализа русского синтаксиса (MARS).

1. Общая стратегия решения проблемы

В настоящем исследовании рассматривается проблема поиска синтаксического хозяина предложных групп (ПГ) в обычном художественном тексте (не записи и не имитации устной речи) и в словарных определениях энциклопедических словарей.

Поиск хозяина предлога (р) — одна из самых трудных задач при построении синтаксических связей.

Для успешного ее решения необходимо определить особенности поведения ПГ в разных контекстах и принципиальную схему анализа.

Проблема рассматривается в рамках подхода, принятого в системе MARS, разрабатываемой в настоящее время в РГГУ. Ее специфика заключается в том, что собственно моделированию структуры предикативных ситуаций, составляющих предложение (S), предшествует сегментация — определение зон его линейной структуры, где могут находиться члены этих предикативных ситуаций.

Система состоит из 6 процедурно независимых модулей [2].

1) Постморфология — несловарные проблемы морфанализа: обработка имен собственных и названий, окказиональных аббревиатур, числительных [3];
2) Разрешение частичной омонимии — совпадения форм разных частей речи [4];
3) Предсегментация — построение **проективных фрагментов** (ПФ) именных и предложных групп (ИГ и ПГ) [5]. ПФ — часть S, границами которой служат связанные подчинительной связью слова. Например, существительное (N) и его слуга — необособленное согласованное определение с вершиной A*– прилагательным, причастием, местоименным прилагательным или порядковым числительным (например, ***лежащий на ложе в грозовом полумраке прокуратор —*** ПФ из прозы М. Булгакова (далее Б), границы задает связь ***прокуратор*** R ***лежащий***); существительное и его слуга-существительное в роли необособленного приложения (***командир алы — сириец*** (Б)); предлог (р) и его слуга-существительное (или его синтаксический субститут) (***на зеленой, никогда не езженной мостовой*** (М)), конструкции с числительными, сложные сказуемые и др., т. е. связи, определяющие единицы линейной структуры, необходимые для сегментации.
4) Сегментация — построение **сегментов** [6], то есть простых-главных S, придаточных S, деепричастных оборотов, обособленных согласованных определений, обособленных приложений, уточняющих и всех других оборотов, обособляемых по правилам русской пунктуации.
5) Моделирование структуры синтагматических связей внутри сегментов.
6) Построение связей сегментов [7].

Поиск синтаксического хозяина ПГ происходит на 5-ом этапе

Сегментация и построение проективных фрагментов чрезвычайно сужают зоны, в пределах которых ведется поиск связей слов, что на практике облегчает решение задачи.

[1] Доклад подготовлен при частичной поддержке РФФИ — грант № 09-06-00275-а

Таким образом, при поиске хозяина ПГ границами поиска хозяина являются границы ПФ, (если мы находимся внутри ПФ), границы сегмента, если мы находимся вне ПФ (при этом слова внутри ПФ из рассмотрения исключаются) и, если ПГ находится внутри сегмента между двумя сочиненными словами, зону поиска определяет свойство сочинительной проективности [8].

2. Синтаксические особенности функционирования предложных групп

2.1. **p** по способности выступать слугами тех или иных частей речи делятся на три класса: 1) **p**, способные быть слугами только глагола (**V**), 2) **p**, которые могут выступать и слугами V и слугами прилагательного (A), и, наконец, 3) **p**, которые могут быть слугами и V, и A, и существительного **(N)** [1]. Естественно, что наибольшие трудности возникают при поиске хозяина предлогов третьего класса.

При этом ПГ, во-первых, может замещать некоторую валентность предикативной вершины сегмента (или ПФ) или выступать в роли обстоятельства этой вершины (т. е. быть актантом или сирконстантом вершины). В этих случаях ПГ подчинена вершине.

Во-вторых, ПГ может быть слугою актантов и сирконстантов или их слуг, чаще всего представленных N.

Последняя группа случаев, когда ПГ подчинена не вершине сегмента или ПФ, распадается на два вида ситуаций.

2.1.1 Хозяевами ПГ являются некоторые слова — не вершины ПФ и сегментов — и эти слова имеют некоторое предложное управление (предложную валентность). Если ПГ с вершиной — этим **p** находится в определенной позиции относительно слова — потенциального хозяина этого **p**, то оно и будет скорее всего ему подчинено. Например, ***Влияние этого человека на ее сына ее пугало.*** Возможно совпадение предложного управления вершины и некоторого слова в сегменте. При этом у каждого из них может быть свой слуга, например: ***Его влияние на сына действовало на всех.*** Но такая ситуация может и породить неоднозначность, например: ***Негативное влияние этого человека на ее сына действовало очень заметно.***

2.1.2. Существуют структуры, в которых ПГ, независимо от возможностей вершины и других слов управлять этим **p**, выступают слугою некоторого слова, для поиска которого оказывается необходимым учитывать и слугу **p** и управляющее им слово, Например: ***Она привезла туфли из кожи. Нужна трубочка длиной со спичку. Этим занимается комитет по правам человека. Он увидел человека в синем плаще.*** Для определения синтаксических хозяев ПГ= ***со спичку \ по правам (человека)\ на концерт \ в синем плаще \ из кожи*** необходимо определить условия, по которым ПГ в этих контекстах не могут быть объявлены слугою сказуемого и не могут заполнять валентность другого слова.

Или же, например, в каждом из предложений типа ***Она из Питера привезла туфли из кожи. Положи билет на концерт на стол.*** мы находим по две ПГ с одним и тем же предлогом: ***из Питера, из кожи; на концерт, на стол.*** При этом надо задать условия, определяющие, какая именно ПГ является определительной, а какая — подчинена предикату: ***Она привезла туфли из Питера. Она только вчера привезла из мастерской туфли из Питера. Она привезла туфли из кожи. Она вынула туфли из Питера из шкафа.***

Подобные ситуации естественно объединяются в классы.

Чтобы в подобных случаях правильно определить хозяина ПГ или не потерять синтаксическую неоднозначность, задаются описания контекстных ситуаций таких классов с указанием порядка следования ИГ и ПГ в тексте и семантических классов или списков хозяев и слуг предлогов.

Например, класс структур типа ***туфли из кожи, нож из стали, чашка из фарфора*** и др. [2] описывается как [ИГ с вершиной N ∈ семантическому классу (СК) ‘предмет неодушевленный’ + ***из*** со слугою N ∈ СК 'вещество, материал'].

Процедуры анализа строятся с учетом того, что и в этих ситуациях могут возникать синтаксические неоднозначности. Например, ***Эта фабрика шьет сумки из крокодиловой кожи*** синтаксически неоднозначно из-за управления ***шить***[3].

3. Встречаемость разных ситуаций в текстах. Статистические наблюдения

Сужение зоны поиска хозяина в результате построения ПФ на этапе, предшествующем моделированию внутренней структуры сегмента и на этапе сегментации, приводит к тому, что часто в потенциально альтернативных случаях при поиске хозяина реально альтернативы не возникает.

Ниже приводятся результаты небольшого статистического обзора распределения типов контекст-

---

[2] Заметим, что приведенные примеры приобретают бóльшую языковую достоверность при наличии у слуги предлога определения, например, ***туфли из мягчайшей кожи, нож из дамасской стали, чашка из саксонского фарфора.***

[3] Не очевидно, что подобную неоднозначность стоит вводить, т. е. учитывать содержательное различие этих интерпретаций.

ных ситуаций, возникающих при поиске хозяина ПГ, для художественной прозы и статей энциклопедического словаря (ЭС). По полученным результатам можно судить, насколько универсальна проблема поиска хозяина ПГ для текстов столь разных жанров, в частности, насколько эта проблема при анализе статей ЭС отличается от аналогичной задачи в обычных художественных текстах, на которые в первую очередь ориентирована система MARS.

Результаты представлены в Табл.1 и Табл.2.

Для анализа художественного текста были выбраны произведения с почти 100-летним разрывом во времени и совершенно разные по литературным критериям. Обращает на себя внимание близость результатов.

В Табл.1 рассмотрены следующие типы контекстных ситуаций.

1. В сегменте или в ПФ есть вершина, которая и является хозяином ПГ. *Это был повергнутый **в пыль** хаос иудейский*.(М) В зоне поиска хозяина, которая определяется границами ПФ *повергнутый **в пыль** хаос,* предикативная вершина — *повергнутый*.
2. Хозяин ПГ определяется тем, что в зоне поиска в определенной позиции по отношению к предлогу есть N, имеющее «преимущественное право» управление данным р: [***отверстие, специалист, участие, статус, ситуация…***] **R *в*$_{предл}$**; [***вступление, выборы, кандидат…***] **R *в*$_{вин.}$** [***взгляд, ответ, право, цена…***R ***на***$_{вин}$]; [***борьба, диалог, дискуссия, договор, отношения, переговоры, связь…*** R ***с***$_{тв}$] и т. п.

Например: *Я всегда смутно чувствовал особенное **значение** Финляндии для петербуржца* и… То, что хозяином ПГ *для петербуржца* является не вершина сегмента *чувствовал*, определятся способностью слова ***значение*** управлять р=для$_{род}$.

Для подобных случаев невозможно обойтись без информации об управлении конкретными предлогами, например: ***песенка о***$_{предл}$, ***жертва за***$_{вин}$, ***игра в***$_{вин}$, ***ударение на***$_{тв}$, ***специалист по***$_{дат}$, ***сравнение с***$_{тв}$ и др.

3. В зоне поиска хозяина в результате построения ПФ и сегментов существует единственный претендент на роль хозяина.
   а) *Семен Афанасьевич Венгеров, родственник мой **по матери**,…*(М) В зоне поиска — обособленном приложении *родственник мой **по матери** —* есть единственный претендент на роль хозяина ПГ.
   б) *…не менее интересно и значительно, чем майский парад **на Марсовом поле**.*(М)
   в) *Там, на Торговых, попадаются еврейские вывески с быком и коровой, женщины с выбивающимися из-под косынки накладными волосами и семенящие …*(М)

Хозяин ПГ = *из-под косынки* однозначно определяется в результате построения ПФ ИГ *выбивающимися из-под косынки накладными волосами*, а в качестве хозяина ПГ = *с … волосами — женщины* и хозяина ПГ = *с быком и коровой — вывески* однозначно определяются по условию сочинительной проективности.в результате построения в ходе сегментации сочинительной связи слов *вывески* и *женщины*

4. Хозяина определяет вхождение ПГ в специфическую структуру, для определения которой нужно учитывать и порядок слов, и семантику хозяев, и семантику ПГ. Например: *Яростно прижимает к груди **книгу** с надписью…* Или *И вдруг два **господина** в цилиндрах, прекрасно одетые, лоснящиеся богатством,…*(М)
5. Есть альтернатива и есть синтаксическая неоднозначность.

Например: а) ***Винодел** с дочками **уехал** к бабушке.* (М) Возможны две интерпретации: ***винодел*** R *с (дочками*) и ***уехал*** R *с*. б) …, *но настоящим **адом** для большинства неловких, не слишком здоровых и нервических детей **был** ручной труд.*(М) Возможны две интерпретации: ***ад*** R *для* и ***был*** **R** *для*.

Заметим, что в прозе Пелевина существенно меньше ситуаций 3-его типа: у Пелевина практически нет сегментов с именными вершинами, в частности, назывных предложений, которые так любит Мандельштам, и сегментов-приложений.

Жанр текста	Всего ПГ в выборке	1. VRp	2. N^{спис}Rp	N^{неглаг} Rp		
				3. Хозяин — не глагол и в зоне поиска хозяина ПГ нет альтернативы	4. В зоне поиска хозяина ПГ есть альтернатива, но нет синт. неоднозначн.	5. В зоне поиска хозяина ПГ есть альтернатива и есть синт. неоднозначн.
Худ. проза — Шум времени, О. Мандельштам	562 100 %	403 72 %	30 6 %	48 8 %	69 12 %	13 2 %
Худ. проза — Поколение «П», В. Пелевин	269 100 %	210 78 %	22 8 %	3 1 %	30 11 %	4 2 %

Табл. 1. Результаты статистического обзора ситуаций управления предлогами в «Шуме времени» О. Мандельштама (далее М) и «Поколении “П”» В. Пелевина (далее П)

Жанр текста	Всего ПГ в выборке	1а. VRp	1б. НМП = N_{nom} R p	2. $N^{спис}$ R p	$N^{неглаг}$ Rp		
					3. Хозяин — не глагол и в зоне поиска хозяина ПГ нет альтернативы	4. В зоне поиска хозяина ПГ есть альтернатива, но нет синт. неоднозначн.	5. В зоне поиска хозяина ПГ есть альтернатива и есть синт. неоднозначн.
ЭС	266 100 %	104 39 %	92 34 %	45 17 %	7 3 %	13 5 %	5 2 %

Табл. 2. Статистика распределения ситуаций управления ПГ в статьях ЭС

В данные таблиц не включены прямая речь и сегменты с нулевым V-предикатом, такие, как: …*махнул рукой, как бы* ***(*будучи)*** *в обиде на такую деловитость.; Да у него самого под столом* ***(*есть)*** *такие надписи. У всякого бренда* ***(*есть)*** *своя легенда.* (П).

Для определений словарных статей «Советского энциклопедического словаря» (ЭС), который после 1991 г. выходил под названием «Новый энциклопедический словарь», был проведен анализ специфики синтаксической и линейной структуры статей.

Синтаксическая структура определений, как правило, начинается с сегмента с именным сказуемым $N_{им}$. и осложнена чаще всего вставлением обособленных согласованных определений разного рода и придаточных с подчинительным союзом ***который***.

Несмотря на тяготение к именным структурам, случаи глагольного управления предлогом (VRp) в ЭС составляют почти половину. Во-первых, они входят в придаточные определительные с глагольными сказуемыми и деепричастные обороты, например: *Абсентеизм земледельческий — форма землевладения, при которой собственник земли,* ***не участвуя*** *в процессе производства,* ***получает*** *денежный доход в виде ренты или прибыли.* И, во-вторых, VRp появляются в обособленных и необособленных согласованных определениях, которые в именных структурах компенсируют отсутствие морфологического сказуемого. Например: *Абрис — схематический план, сделанный от руки, с обозначением данных полевых измерений, необходимых для построения точного плана или профиля*.

В ЭС часто встречаются необычные для нормальной повествовательной структуры конструкции: первое именное сказуемое — $N_{им}$, , и оно входит в цепочку сочиненных сказуемых, которые, начиная со второго, уже представлены обычными морфологическими сказуемыми. (Надо отметить и необычную при этом пунктуацию: зоны влияния этих сказуемых разделены точкой с запятой, что практически не встречается в других жанрах). Например: *Абонентское телеграфирование —* ***разновидность*** *телеграфной связи;* ***служит*** *для непосредственного двустороннего обмена телеграфными сообщениями между абонентами.*

Таким образом, в ЭС грамматически возможны все те же варианты соединений сегментов в одном S, что и в художественном тексте, а значит, требуются те же процедуры сегментации.

4. Словарь исключений — «частных случаев» (ЧС)

Словарь ЧС описывает контекстные условия, определяющие безусловного хозяина ПГ с точностью до порядка слов и семантических классов (или списка лексем) хозяев и\или слуг конкретных предлогов.

Использование словаря ЧС удобно: оно позволяет, не меняя словаря основ, легко пополнять набор типовых структур, представляющих как большие классы слов, так и уникальные словосочетания.

К подобным конструкциям относятся хорошо известные комитативные группы, где слугами являются ПГ с вершиной ***с***, управляющей $N_{тв}$.(***мы с приятелем, дом с мезонином, машина с желтыми фарами*** и т. д.)

В словаре ЧС представлены ситуации для двух типов комитативных групп — существительных неодушевленных и одушевленных:

- ***тележка с двигателем\ вагон с двигателем \ автомобиль с кузовом \ форма с рельефными печатающими элементами\ трансформатор с одной обмоткой*** и т. п.
- ***инженер И. Ф. Иванов с другом\ инженер нашего завода Ваня Петров со своей младшей сестрой \ Иван с другом\мы с Ваней*** и т. п.

В некоторых конструкциях и хозяин **р** и его слуга — большие группы слов. Иногда их удается задать семантическими классами (СК). Например:

- N^j …+ p=***для*** R N^v (N^v — глагольное N), где N^j є СК 'устройство, приспособление' (***аппарат для проведения; прибор для автоматического удержания; экскаватор для переукладки*** и др.(ЭС)).
- [N^j — ´параметр + ***с*** R $N_{вин}$, где $N_{вин}$ — ´предмет неодуш.´] (например, ***толщиной \длиной со спичку***).
- [N^v (+$ИГ^к_{род}$.)(+ $ИГ^{к+1}_{род}$) + ***с помощью$_{род}$\при помощи*** R $N^{к+2}$] где $N^{к+2}$ ´прибор, аппарат´

(например, ***опыление … с помощью пульверизатора \ вертолета***)

- [N^j + р = ***в виде*** R $N^к$] где N^j и $N^к$ — ´предмет неодуш.´ (например, ***застежка в виде броши***)

Значимы в силу высокой частотности две конструкции с предлогами ***в***$_{предл}$ и ***без*** с большими списками и хозяев и слуг, частые в текстах любых жанров:

- N^{j*} +[ПГ= *в* R Nк*] (например, ***старичок в соломенной шляпе \ дама в бальном платье \ подсудимый в наручниках***)
- Nj* + ***без*** R Nк* (например, ***старичок без пиджака***)

N^{j*} — нарицательные имена человека, групп людей или изображений человека, различающиеся по полу и возрасту: ***человек, мужчина, старик, старичок, старуха, старушка, дядя, тетя*** …; по роду деятельности: ***король, царь, премьер (-министр), председатель, конгрессмен, заседатель, судья, терапевт, психиатр, отоларинголог…***; изображения человека: ***фигура, манекен, статуя***, …; собирательные одушевленные: ***толпа, свита, солдатня, колонна (демонстрантов)…***

N^{k*} — предметы, которые могут быть надеты на человека, служить одеждой, украшением и т. д.: ***пальто, плащ, горжетка, кольчуга…; шляпа, кепка, платок, шлем, каска….; галстук, шарф, …; бусы, кольцо, перстень, браслет, наручники, ошейник, очки, пенсне…***

В некоторых ситуациях для небольшого набора ПГ существует большой список хозяев, который не всегда можно задать семантическим классом.

- [N^j=N^v+ (ИГ^{к}род.(ИГ$^{к+1}_{род}$)) + в$_{предл}$ R ИГ$^{к+2}$] где вершина ИГ$^{к+2}$ = ***цель*** (например, ***конструирование и постройка моделей автомобилей в технических или спортивных целях*** (ЭС)).
- [N^j — имя родственника (***дядя, тетя, дедушка, бабушка…***) (+Nj+к — имя родственника в Род)]+ [[***со стороны \ по линии*** R ***матери \ отца***] **v** [***с материнской \ отцовской стороны***]]
- [N^j ***…*** + (Ngen)+ ((ПГ)+ в N^{num}***несколько\много +этажа\этажей***] , где одно из значений Nj — строение: ***строение \ дом \ башня \ небоскреб \ магазин \ театр…***
- **[**N^j **…** ПГ= ***по правам + человека***]**:** где между N^j и ***по*** м. б. только ИГ в Род. п., ПГ, D, частицы.
- N^j = ***евросуд, конвенция, комитет, документы, госбюро, центр, обзор, законодательство, образование, просвещение, библиотека, уполномоченный, линия, суд, комиссар…***
- [$N^J_{одуш\ собират}$ + ***во главе с*** R $N^к_{Тв.}$] где N^J — организация или собирательное одушевленное, а $N^к$ — одушевленной нарицательное или собственное (например, ***фирма \ школа \ группа \ класс во главе с Иваном Ивановичем***).

В настоящее время словарь ЧС этого алгоритма насчитывает 35 линейных конфигураций[4]. Список ЧС открыт. Многие структуры для их введения в словарь ЧС требуют наблюдений за их поведением в тексте.

5. Алгоритмическое решение задачи

Построен алгоритм, реализующий следующую принципиальную схему анализа.

Нормой в алгоритме считается 1-ая ситуация: ПГ объявляется слугою вершины сегмента, которая в простом-главном и придаточном S м. б. выражена и неморфологическим сказуемым, в частности — $N_{им}$,. Если ПГ вложена в ПФ, то ее хозяином объявляется предикативное слово — вершина ситуации, границы которой заданы ПФ.

Чтобы объявить вершину сегмента или вершину предикативной ситуации ПФ хозяином, необходимо удостовериться, что в зоне влияния этой вершины нет ситуаций — исключений из этого правила, представленных в Табл.1 столбцами 2 и 4, когда управление ПГ «перехватывает» слово — не вершина.

Существенно, что в обычных текстах с цепочками именных и предложных групп вершин — $N_{им}$ часто появляется в простых назывных S. В ЭС $N_{им}$ обычно в выступает в роли именного сказуемого, причем в силу специфики задач текста это сказуемое распространено и косвенными падежами существительных и предложными группами. Например: *Аэрозольный генератор — **машина** для образования термомеханическим способом аэрозолей пестицидов и их распыления в целях уничтожения вредителей сельскохозяйственных культур, лесных и плодовых насаждений, а также для обеззараживания сельскохозяйственных хранилищ и животноводческих помещений.*

При поиске хозяина ПГ в таких сегментах с именным сказуемым приходится решать те же проблемы, что и в сегментах с глагольными вершинами, т. е. ситуации 2-ого, 3-его и 4-ого столбцов играют ту же роль, что в сегментах с глагольной вершиной и используется тот же лингвистический базис.

Соответственно, качество анализа определяется

- точностью отбора сегментов, для которых в результате сегментации не возникает альтернативы (ситуации 3-его столбца Табл.2),
- полнотой словаря специфических конструкций со слугами-ПГ (4-ый столбец) и

[4] Программная версия алгоритма показывает хорошие результаты применения предлагаемой стратегии при обработке тестовых примеров, и на следующем этапе отладки системы, когда список ЧС достигнет определенной степени полноты, предполагаются эксперименты, которые позволят определить количественные оценки работы алгоритма.

- полнотой списков слов — хозяев конкретных предлогов (для ситуаций 2-ого столбца).

Первый шаг анализа — поиск ситуаций, когда хозяина можно найти без анализа управления. Для этого алгоритм определения безальтернативного управления ПГ ищет признаки контекстов, в которых хозяин ПГ определяется однозначно линейной структурой сегмента.

Второй шаг — поиск случаев-исключений. Если не найдена безальтернативная ситуация, алгоритм обращается к словарю ЧС и проверяет, не соответствует ли контекст анализируемой ПГ одной их структур этого словаря.

Третий шаг. Если нет ситуации безальтернативного управления и нет ситуации словаря ЧС, проверяется, что в сегменте или ПФ нет лексем, способных управлять конкретным предлогом, которые могут «перехватить» управление ПГ у вершины. В MARS в словаре основ управление предлогом лексически не конкретизировано, поэтому в специальной таблице для каждого **p**, управляющего определенным падежом и способного быть слугою и V и N, задается список N. В настоящее время такие списки заданы для 34 предлогов и суммарно насчитывают 428 существительных. При появлении в определенной позиции по отношению к ПГ такого N, если нет конкурирующей валентности, мы отдаем предпочтение этому N. Если же эта ПГ — единственный вариант заполнения одинаковых валентностей двух слов, фиксируется неоднозначность.

6. Заключение

Предложена и алгоритмически реализована универсальная стратегия поиска хозяина ПГ, опирающаяся на использование зон поиска хозяина **p**, определяемых на этапе сегментации предложения, предшествующем этому поиску.

Анализ ситуаций, возникающих при поиске хозяина ПГ, доказывает возможность построения универсальной грамматики поверхностного синтаксиса для этой зоны автоматического анализа русского предложения.

Успешность работы алгоритма в сегментированном предложении определяется полнотой и точностью описаний, задающих модели предложного управления, в частности насыщением важного ее фрагмента, исчисляющего контекстные ситуации специфического управления предложными группами.

Литература

1. *Мельчук И. А.* Автоматический синтаксический анализ. Т. 1. — Новосибирск: Ред.-изд. отдел Сибирского отделения АН СССР, 1964.
2. *Кобзарева Т. Ю.* Иерархия задач поверхностно-синтаксического анализа русского предложения // НТИ, Сер.2, №1, 2007 — С 23–35.
3. *Кобзарева Т. Ю.* Морфанализ in vivo // Труды Международной конференции Диалог'2004, — М.: Наука, 2004 — С. 286–291.
4. *Кобзарева Т. Ю., Афанасьев Р. Н.* Универсальный модуль предсинтаксического анализа омонимии частей речи в русском языке на основе словаря диагностических ситуаций // Труды международного семинара Диалог'2002, Протвино 2002. Т. 2. С. 258–268.
5. *Кобзарева Т. Ю.* Некоторые свойства линейной структуры именных и предложных групп (Поверхностно-синтаксический анализ русского предложения)// Вестник РГГУ, № 8/07, Серия «Языкознание» (Московский лингвистический журнал № 9/2) — М.: 2007. — С. 113–130.
6. *Кобзарева Т. Ю.* Принципы сегментационного анализа русского предложения. Московский лингвистический журнал // М.: 2004. Т. 8. №1 — С. 31–80
7. *Кобзарева Т. Ю.* Построение графа связей сегментов (поверхностно-синтаксический анализ русского предложения) // Компьютерная лингвистика и интеллектуальные технологии. Труды Международной конференции Диалог, М, Наука, 2008 — С. 192–198.
8. *Кобзарева Т. Ю.* Рекурсивность и проективность сочинительных связей в русском тексте // Труды Международной конференции Диалог 2006, Бекасово — М.: Наука, 2006. — С. 223–229.

Онтология силовых процессов

Force interactions ontology

Кобозева И. М. (kobozeva@list.ru),
Марушкина А. С. (nastam@rambler.ru)
МГУ им. М. В. Ломоносова

В статье предпринимается попытка показать, как созданная на основе теории Динамики Сил онтология силовых процессов может быть использована для такого семантического представления текста на естественном языке, которое может служить базой для естественно-логического вывода о результатах и дальнейшем развитии описанных в нем ситуаций силового взаимодействия объектов. Семантическое представление текста задается, с одной стороны, в виде сценария, учитывающего основные силовые переходы, а с другой — в виде формализованного лексического описания.

1. Границы предметной области

Разработка коммерческих онтологий на данный момент сводится в основном к представлению четко очерченной и хорошо структурированной предметной области, например, онтологии медицинских препаратов или молекулярных функций, вин и продуктов питания и пр. Отношения между концептами в такой онтологии являются достаточно регулярными, а вопрос о границах предметной области не представляет особой сложности.

В случае создания универсальной онтологии, предназначенной для моделирования смысла естественно-языковых текстов по любой тематике на том уровне понимания, который отражает представления о мире рядового носителя языка, этот вопрос, безусловно, не является тривиальным, потому что мы имеем дело с бесконечно сложно организованным объектом моделирования. Во-первых, характер связей между концептами может быть самый разный, во-вторых, в семантике слов семы, соответствующие концептам онтологии, могут обладать разной степенью выраженности.

Цель данной работы — показать вариант анализа текста, описывающего силовые процессы, с использованием фрагмента универсальной онтологии. Организующим звеном для его построения является понятие Динамики сил, введенное в лингвистический обиход Леонардом Талми (см. [6]).

Следуя определению, которое Талми дает категории Динамики сил, обозначим границы предметной области следующим образом: **мы рассматриваем любые концепты, отражающие изменение состояния физических объектов в процессе силового взаимодействия**. Причем силовые процессы могут быть любого типа: это и действие силы, и сила противодействия (resistance to force), преодоление силы противодействия, блокирование силового воздействия (blockage of the expression of force), снятие блокирования и т. д.

Предлагаемая нами онтология в отношении формальной репрезентации семантической структуры концептов и лексических значений близка к универсальной онтологии «Онтологической семантики» С. Ниренбурга и В. Раскина [5], и поэтому может рассматриваться как дополнительный модуль этой онтологии, подключаемый при необходимости более точного «вычисления» смысла описываемых в тексте ситуаций силового взаимодействия объектов. На данном этапе она позволяет проинтерпретировать в аспекте динамики сил только тот пласт лексики (в основном глагольной), представители которого имеют в своем составе сему «сила», соответствующую онтологическому концепту СИЛЫ (FORCE) как физической величины. При этом пока не учитывается вклад граммем вида и залога в смысл языковых описаний ситуаций силового взаимодействия, а также метафорические употребления рассматриваемых единиц, хотя можно предположить, что структура выделенных концептов может помочь определить нетривиальные основания для метафорического переноса.

Особенность онтологии, построенной с помощью категории Динамики Сил, заключается в том, что с ее помощью можно учесть те аспекты означаемого, которые важны для естественно-логического вывода по тексту (иногда их рассматривают как следствия или импликации [2], [4]).

2. Онтология, лексикон и семантическое представление текста

На данном этапе мы выделяем два модуля, задействованных при семантическом анализе текста. Это онтология и лексикон.

Процедура обработки текста представляет собой процесс, протекающий на разных уровнях. В рамках данного исследования разрабатывается подход к построению семантического представления текста без привязки семантического блока к другим блокам обработки текста (морфологическому и синтаксическому парсерам).

Онтология Динамики Сил представляет собой набор концептов, репрезентирующих внутреннюю динамику силовых процессов, закономерности их протекания и позволяющих логически связать несколько примитивных ситуаций в единый сценарий.

Лексикон позволяет учесть дополнительные параметры ситуаций, которые будут влиять на выбор той или иной лексической единицы при описании ситуации.

3. Онтологический концепт и его структура

Основное отличие разрабатываемого нами фрагмента онтологии от других доступных онтологий состоит во внутреннем устройстве концептов и характере отношений между ними. Во многих лексических ресурсах, концептуализация законов силовых взаимодействий отражена номинативно, то есть внутренняя структура соответствующих концептов в полной мере не описана. Это происходит даже в тех случаях, когда создатели ресурса делают успешную попытку моделировать концептуализацию силовых процессов не в научной, а в языковой картине мира и делают это достаточно формально (см. [5], [7], [8], [9]). Как было показано раньше в работе одного из авторов данного исследования (см. [3]), языковая концептуализация силовых взаимодействий может быть представлена в терминах *тенденций* и *соотношений*, приводящих к определенному результату. На основе предложенных Талми концептуальных схем мы выделили 10 типов базовых ситуаций, которые отражают наивное представление о законах физической механики. Таким образом, в нашем распоряжении оказывается 10 онтологических концептов. К сожалению, описание каждого из них достаточно трудоемко, но поскольку для понимания устройства онтологии достаточно объяснить базовые принципы построения концептов, мы этим и ограничимся (для более подробного ознакомления со структурой каждого концепта отсылаем читателя к работе [3]). Каждый выделенный тип силового взаимодействия представлен в виде одной или нескольких схем, структурными элементами которых являются:

- Участники ситуации: антагонист (Ant) — воздействующий, и агонист (Ago) — претерпевающий воздействие
- Внутренняя тенденция участников: импульс к действию (toward action) — тенденция к покою (toward rest)
- Соотношение сил: более сильный (stronger entity) — менее сильный (weaker entity)
- Результат взаимодействия: действие (action)[1] / движение (motion) — покой (rest)

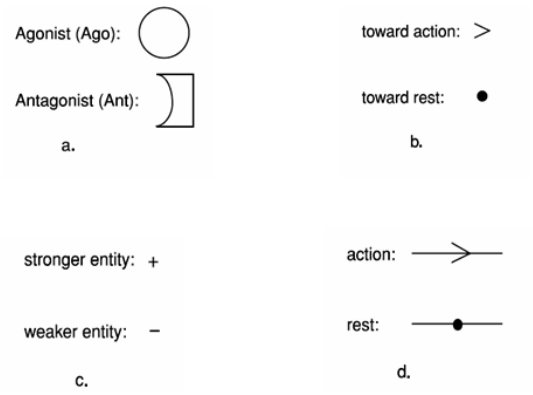


Рис. 1

В виде сценарной схемы силовых процессов представим следующее предложение в его «силовой» интерпретации (другие возможные для него интерпретации в рамках данного исследования игнорируются):

(87) Дверь не открывается

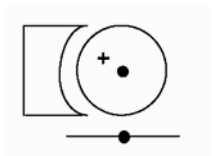


Рис. 2

Схема показывает, что на дверь воздействует некоторая сила, которой недостаточно, чтобы дверь открыть, поэтому результатом взаимодействия оказывается состояние покоя.

Не все концепты, однако, описываются одной схемой — некоторые имеют более сложную внутреннюю логическую структуру. Для примера рас-

[1] В рамках данной статьи не будем разграничивать понятия *движение (motion)* и *деформация (deform)*, состоянию *покой (rest)* будем противопоставлять *изменение (action)*, под которым следует понимать дизъюнкцию (*движение ^ деформация)*.

смотрим глагол со значением импульсного воздействия *толкать*. Успешность воздействия неопределена, а значит, и сценарное описание данного силового процесса не может быть однозначным. По этой причине мы будем описывать соответствующий ему концепт через дизъюнкцию, которая показывает, что движение (деформация) может не последовать, если силы воздействующего для этого не достаточно:

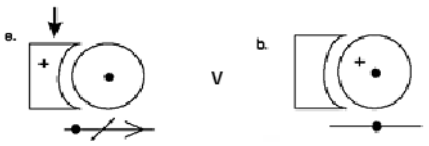


Рис. 3

Структурные элементы схем формально представлены в виде слотов. Набор слотов фиксирован для каждого из концептов и выглядеть он будет следующим образом:

```
I.    1. Участники ситуации:
         ANT (SEM (PHYSICAL_OBJECT))
         AGO (SEM (PHYSICAL_OBJECT))

      2. Внутренняя силовая тенденция:
         (HAS_TENDENCY (SEM ({ACTION, REST, MOTION\DEFORM})))

II.   3. Соотношение сил:
         FORCE_BALANCE (SEM ({UNKNOWN, LETTING, STRONGER (SEM (ANT)),
                              STRONGER (SEM(AGO))})))

III.  4. Начальное состояние Агониста:
         FD_INITIAL (SEM ({REST, MOTION\DEFORM}))

      5. Результат взаимодействия:
         FD_RESULT (SEM ({UNKNOWN, ACTION, MOTION\DEFORM, REST}))
```

Необходимо оговориться, что выбор фасетов — наследие терминологии принятой в [5]. В фигурных скобках через запятую перечислены концепты, составляющие область значений слота. За первыми двумя слотами на данном этапе закреплено значение PHYSICAL_OBJECT (физический объект) — метафоры мы пока не рассматриваем.

По законам Динамики Сил Агонист и Антагонист должны стремиться либо к изменению (движение, деформация), либо к покою. Это отражено в слоте HAS-TENDENCY. При этом тенденции Антагониста и Агониста в рамках одной схемы не могут совпадать, поэтому достаточно задать только тенденцию Агониста. Значение слота HAS-TENDENCY для Антагониста будет определено как обратное. Для фиксации этой закономерности в онтологии используем слот OPPOSITE.

Для обозначения соотношения сил между основными участниками внутри концепта вводим четыре значения слота FORCE_BALANCE: STRONGER (SEM (ANT)), STRONGER (SEM (AGO)), LETTING, UNKNOWN. Значение LETTING соответствует снятию силового воздействия, которое Антагонист оказывает на Агониста, а значение UNKNOWN показывает, что в данном концепте соотношение сил не определено.

Значения слота FD_RESULT (результат взаимодействия) показывают конечное состояние Агониста (изменение или покой).

Осталось фиксировать только динамический контур ситуации — постоянный или переменный (то есть, меняется ли начальное состояние Агониста или нет). Для этого целесообразно ввести еще один слот, который будет отражать начальное состояние Агониста, а вместе с ним и системы объектов в целом. Такой слот назовем FD_INITIAL**.** Если его значение одновременно совпадает со значением HAS-TENDENCY для Агониста и не совпадает со значением слота FD_RESULT, мы имеем дело с ситуацией силового взаимодействия с **переменным** динамическим контуром.

4. Создание лексикона и формата семантического представления анализируемого текста

Теперь на основе описанных выше концептов можно начать разработку лексических описаний для создания словаря. Предложенная структура лек-

сического входа разработана для описания лексических единиц русского языка с семантикой силового взаимодействия.

4.1. Структура шаблона для лексического описания

Как было отмечено ранее, семантическое представление текста имеет двойной формат: с одной стороны, это сценарное описание с привлечением концептов онтологии, с другой — описание в формате словаря. В этом разделе рассмотрим шаблон лексического входа для единиц с силовым значением.

В первую очередь рассмотрим свойства Антагониста (воздействующего). Пока здесь выделены только два слота: это *одушевленность / неодушевленность* и характеристика *лицо / не-лицо*. Важно сразу сказать, что эти два параметра отвечают за разные области. *Одушевленность* во многом определяет тип Агониста (такие его характеристики, как, например, манипулируемость), а также распределение сил между Агонистом и Антагонистом. Эта характеристика описывает физическую сферу. *Лицо / нелицо* — параметр, главным образом определяющий импликации, которые станут основой для дальнейшего семантического расширения. С ним неразрывно связано понятие цели действия (ср. глаголы *бросить* и *выбросить*). От того, является ли Антагонист лицом, могут также зависеть и такие существенные вещи, как выбор предлога. Например, *тащить* что-либо можно *за* собой и *на* себе, только последний предлог предпочтительнее именно в том случае, когда Антагонист — лицо.

В работе [1] также отмечается, что идея агенса напрямую связывается с идеей воли и цели. При этом акцентируется семантическая сопряженность с этой идеей значений граммемы несовершенного вида и в качестве подтверждения этому наблюдению приводятся деструктивные глаголы (*рвать, разбивать, ломать* и др.), которые не имеют актуально-длительного значения, когда человек выступает не в роли агенса, а в роли «невольной причины» разрушения: ср.

(88) Его толкают, и он разбивает окно,

где деструктивное действие не намерннное, и соответствующие глаголы употреблены нарративно, и

(89) Посмотри-ка, они уже и витрины разбивают,

где агенс уже действует целенаправленно, и реализуется актуально-длительное значение глагола НСВ (примеры взяты из [1]).

В общую концептуальную структуру также вводится описание физических и топологических свойств Антагониста.

Описание Агониста (претерпевающего воздействие) устроено похожим образом.

Остальные слоты являются факультативными. Это значит, что их наличие в той или иной степени зависит от онтологического класса. Так, например, не всегда необходимо исчерпывающее описание физических параметров Антагониста (таких как размер и вес). Для онтологических концептов, которые фиксируют соотношение сил между Антагонистом и Агонистом, к которым, например, относятся глаголы *двигать, держать*, описание размера и веса обычно необязательно.

Еще один факультативный параметр — форма. Форма Агониста играет решающую роль при выборе лексической единицы в ограниченном числе случаев. Самый яркий пример — предикат *катить*, который в общем случае предполагает, что Агонист круглый / шарообразный.

Далее задается описание свойств материала, из которого сделан Агонист (в случае, если он неодушевленный). Этот раздел лексического описания важен для тех ситуаций, которые потенциально влекут за собой деформацию Агониста. В других случаях эти параметры также можно опустить.

Для ряда случаев (а именно, для ситуаций с переменным динамическим контуром, где Антагонист препятствует движению Агониста) также важно указывать, с какой скоростью Агонист движется. На данном этапе мы предлагаем три значения для этого слота:

```
VELOCITY (high, medium, low).
```

От типов объектов перейдем к характеристике действия силы. Здесь мы имеем принципиальные различия в описании для концептов деформации и концептов движения.

Мы выделяем несколько направлений приложения силы: это направления вверх-вниз, вправо-влево, а также четыре диагональных направления. Вектор приложения силы и наклон поверхности, по которой происходит движение, взаимозависимые величины.

Как мы уже говорили, внутренний семантический потенциал предикатов силового взаимодействия может быть разным: неодинаково распределение значимости между признаками, характеризующими ситуацию, значение одного и того же признака в одних случаях выводится из контекста, в других — оно будет фиксированным. Так, на примере глаголов сопротивления хорошо видно, что в семантической структуре этих глаголов присутствует дополнительный признак «точка зрения» или внутренний фокус эмпатии. Возьмем пару глаголов *опираться* vs *подпирать*. Практически идентичная ситуация дается сначала с точки зрения Агониста, потом — с точки зрения Антагониста. Это явление отражено в слоте `FOCUS_VALUE`.

Остальные три слота, описывающие природу приложения силы, универсальны. Интенсивность приложения силы определяется по шкале от 0 до 1. От того, какое значение получает данный слот в том или ином случае, может зависеть выбор лексемы (ср. *стукнуть — шарахнуть*).

Сила может действовать на Антагониста постоянно или импульсно (ср. *тащить vs. толкать*). Эта характеристика отражена в бинарном слоте APPLICATION (permanent, impulse).

Также выделен слот CONTACT_PATH, с помощью которого кодируется тип контакта Агониста с поверхностью, по которой происходит движение — может иметь место либо касание только одной стороной (single_side), как для глаголов *двигать*, *тащить*, либо касание по всей поверхности Агониста (comprehensive), как для глаголов *катить, перекатывать*. Следующий слот заполняется только при значении предыдущего — comprehensive, то есть контакт по всей поверхности Агониста. Это слот PROMPTNESS, или количество оборотов, совершенных Антагонистом, каковое представляется в виде численного значения.

Наконец, последний слот в этом разделе — CONTACT_ANT — описывает опосредованность контакта между Антагонистом и Агонистом, т. е. специфицирует, подразумевается ли в данной ситуации дополнительный инструмент воздействия. Его значения: direct (непосредственный) и instr (опосредованный). При выборе второго значения необходимо указать, какой именно инструмент использован при опосредованном контакте. Для этого существует слот INSTR_TYPE, областью значения которого по умолчанию является множество объектов, являющихся артифактами.

В последней части описания описано взаимодействие Агониста с окружающей средой (чаще всего — поверхностью, по которой происходит или из-за физических свойств которой не происходит движение). Сюда же как внешний признак попадет траектория движения Агониста.

Структура лексического описания

```
[1] Participant Type
[1.1] Antagonist
        [Ant_type ANIM(anim / inanim)
                    PERS (pers / non-pers)
                    VOLITIONAL (+,-) {намеренно/ненамеренно}]
                    PHYSICAL_TYPE
[1.2] Agonist
        [Ago_type
                   ANIM(anim / inanim)
                   PHYSICAL_TYPE
                   SHAPE(shape)
                   SIZE (0<>1)
                   WEIGHT (0<>1)
                   MATERIAL (x, (
                          ELASTICITY (0<>1))
                          SOLIDITY (0<>1))
                   ]
                   {В этом разделе в описание попадает ограниченное количество
                    параметров: акцент может делаться как на материале (одной
                    или нескольких характеристиках), так и на форме}
[2] Focus
      [Focus_Value (Ant, Ago)]

[3] Force Application
        [Force_Application
                   VECTOR (towards_Ant, from_Ant, down, up, towards_Ant_down,
                           from_Ant_down, towards_Ant_up, from_Ant_up)
                   INTENSITY (0<>1)
                   APPLICATION (permanent, impulse)
                   CONTACT_ANT (direct, instr)
                   INSTR_TYPE (artifact)
                   CONTACT_PATH (comprehensive, single-side) — {контакт
                                с поверхностью: по всей поверхности агониста
                                или точка соприкосновения с пов-тью фиксирована}
```

```
                    PROMPTNESS(n >= 0){число оборотов}]
[4] Environment
[4.1] Ago_trajectory
        [Ago_trajectory
                    TRAJECTORY (default {= vector value}, circled, wigwag{туда-
                    сюда}, oscillation {вращения}, stop)
                    TRAJECTORY_LENGTH (n >=0)]
                    VELOCITY (high, medium, low, n >=0){скорость движения Агониста}]
[4.2] Environment
        [Path/Environment
                    ROUGHNESS(0<>1){шероховатость}
                    INSTABILITY (+,-) {нестабильность}
                    BIAS (towards_Ant_down (0<>1), from_Ant_down (0<>1),
                          towards_Ant_up (0<>1), from_Ant_up) (0<>1)]
```

4.2. Пример построения семантического представления текста

Попробуем представить общий вариант анализа входящего письменного текста.

Для простоты возьмем текст задачи из сборника задач по физике для 7–9 классов. Употребление лексических единиц с силовой семантикой здесь, как правило, не противоречит обыденному употреблению, а каждый из текстов представляет собой некоторую, обычно завершенную, цепь сменяющих друг друга силовых переходов, для которых удобно показать, как на практике выглядит их последовательный анализ.

Обработка текстов задач существенно отличается от обработки любого другого вида текста только в двух взаимосвязанных аспектах. Во-первых, с дискурсивной точки зрения задачи имеют четкую двухчастную структуру: основную часть и вопрос. Причем в основной части всегда дается развернутый сценарий, а при формулировании вопроса либо дается дополнительная информация для расчетов, либо альтернативные условия протекания силовых процессов. Во-вторых, на уровне конечного семантического представления текста будет заполнено значительно большее количество слотов, чем у обычного текста.

Рассмотрим такой пример:

(90) Мяч, кото́рый урони́ли с не́которой высоты́ H в неподви́жном ли́фте, подпрыгивает на высоту́ h. Изме́нится ли э́та высота́, е́сли лифт равноме́рно дви́жется навстре́чу уро́ненному в нём с той же высоты́ мячу́?

Мы делим текст задачи на две части по границе между предложениями. Сначала система обращается к словарю, с помощью которого определяются лексические единицы со значением силовых процессов. Таким образом обнаруживаются два предиката: *уронили* и *подпрыгивает* и концепты, которым они соответствуют. На основе концептов строим сценарий, который графически будет выглядеть следующим образом (формальное описание концептов, состоящее из слотов, представленных в разделе **3**, полностью соответствует графическому, но является довольно громоздким, поэтому мы не будем приводить его в рамках данной работы):

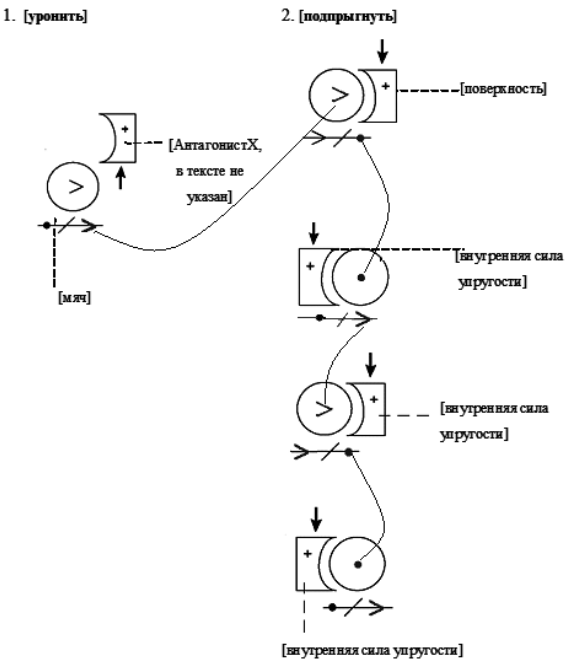


Рис. 4

Глагол *подпрыгивать* (с неодушевленным Агонистом) подразумевает наличие упругого удара, что отражено конъюнкцией сразу четырех концептуальных схем. Причем в двух из них в роли Антагониста выступает физическая сила — внутренняя сила упругости (далее INNER_ELASIC_FORCE).

Теперь на основе лексических входов для этих глаголов начинается построение семантического представления текста задачи, ориентированного на отражение силовой динамики.

«Уронить»
Antagonist [X]
[Ant_type ANIM(anim)
PERS (pers)
VOLIIONAL(-)]

Agonist$_1$ [«мяч»][2]
[Ago_type
ANIM(inanim)
PHYSICAL_TYPE
SHAPE(round)
SIZE (0<>1)
WEIGHT (0<>1)
MATERIAL (x, (
ELASTICITY (0.4<>0.5))
SOLIDITY (0.8<>0.9))
VELOCITY (high, medium, low or n >=0)]
Focus
[Focus_Value (Ant)]
Force Application
[Force_Application
VECTOR (down)
CONTACT_ANT (direct)
]
Ago_trajectory
[Ago_trajectory
TRAJECTORY (default {= vector value})
TRAJECTORY_LENGTH (n >=0)]
Environment [«лифт»]$_E$ {нижний индекс позволяет показать, что *лифт* выступает в качестве окружающей среды в данном сценарии, но в следующем его роль меняется, и поверхность лифта будет играть уже роль Антагониста}
[Path/Environment
ROUGHNESS(0<>1)
INSTABILITY (-)]
«Подпрыгнуть»
Antagonist
[Ant_type$_1$ SURFACE$_E$ Ant_type$_2$ INNER_ELASTIC_FORCE Ant_type$_3$ INNER_ELASTIC_FORCE]
Agonist$_1$ [«мяч»]
Focus
[Focus_Value (Ago)]
Force Application
[Force_Application$_1$
VECTOR$_1$ (down_squeeze)
CONTACT_ANT$_1$ (direct)
Force_Application$_2$
VECTOR$_2$ (up)
CONTACT_ANT$_2$ (direct)
Force_Application$_3$
VECTOR$_3$ (up)
CONTACT_ANT$_3$ (direct)
]
Ago_trajectory$_2$
[TRAJECTORY (stop)

[2] Здесь и далее физические свойства Агониста (кроме свойства VELOCITY), когда он назван в тексте, по умолчанию берутся из других концептов. Так, например, свойства Агониста *мяч* взяты из соответствующего концепта, присутствующего в универсальной онтологии.

```
                    TRAJECTORY_LENGTH (n >=0)]
Ago_trajectory3
                    [TRAJECTORY (vector3 value)
                    TRAJECTORY_LENGTH (H)]
Environment [«лифт»]
        [Path/Environment
                    ROUGHNESS(0)
                    INSTABILITY (-)
                    ]
```

Поскольку в вопросе к задаче изменяется статус стабильности окружающей среды, в альтернативном представлении поменяется значение слота `INSTABILITY` в разделе `Environment` с «–» на «+»

5. Заключение

В данной статье мы попытались показать, как можно построить семантическое представление текста с использованием онтологии силовых процессов. Конечно, в ряде случаев пришлось огрублять и упрощать предложенные описания, кроме того, для качественного анализа текста еще предстоит провести большую работу. Однако основные принципы применения данной онтологии при обработке текста на естественном языке уже достаточно четко прослеживаются. Получаемое формальное сценарное представление, отражающее динамику силовых процессов, обозначаемых глаголами, может служить входом для правил естественно-логического вывода о тенденциях и результатах описанных в тексте ситуаций взаимодействия физических объектов.

Литература

1. *Апресян В. Ю., Апресян Ю. Д., Бабаева Е. Э., Богуславская О. Ю., Иомдин Б. Л., Крылова Т. В., Левонтина И. Б., Санников А. В., Урысон Е. В.; Отв. ред Апресян Ю. Д.* Русская языковая картина мира и системная лексикография // М.: Языки славянских культур, 2006
2. *Кустова Г. И.* Типы производных значений и механизмы языкового расширения. М.: Языки славянской культуры. 2004
3. *Марушкина А. С.* «Наивная механика» в языке и онтологии // Сборник трудов международной конференции Диалог, 2006
4. *Падучева Е. В.* Динамические модели в семантике лексики. М.: Языки славянской культуры, 2004
5. *Nirenburg S., Raskin V.* Ontological semantics // Cambridge.: MIT press, 2004.
6. *Talmy L.* Toward a cognitive semantics. Vol. 1 // Cambridge.: MIT press, 2000.
7. *www.ontologyportal.org* (SUMO)
8. *www.opencyc.org* (CYC)
9. *http://wordnet.princeton.edu* (WordNet)

Альфа и Омега, от А до Я: от исходной метафоры к современному значению

Alpha and Omega, from A to Я: from primary metaphor to contemporary meaning

Козеренко А. Д. (akozerenko@mail.ru)

Институт русского языка им. В. В. Виноградова РАН

В работе проводится анализ идиом русского языка, внутренняя форма которых отсылает к первой и последней буквам различных алфавитов: *альфа и омега, от альфы до омеги, от а до я, от аза до ижицы, от а до зет*. Рассматривается связь значения этих идиом с первоначальной метафорой, с внутренней формой, предлагаются толкования.

Я есмь Альфа и Омега, начало и конец, говорит Господь, Который есть, и был, и грядет, Вседержитель.

Откровение святого Иоанна Богослова, 1.8

0. В настоящей работе рассматривается ряд идиом русского языка, внутренняя форма которых отсылает к первой и последней буквам различных алфавитов: *альфа и омега, от альфы до омеги, от а до я, от аза до ижицы, от а до ижицы, от а до зет*. Будет показано, что эти идиомы описываются четырьмя толкованиями, причем полученные толкования и значения идиом не соответствуют традиционно приводимым в ряде современных фразеологических словарей. Будет выявлено, как описываемые значения идиом кореллируют с внутренней формой идиомы и с категориальной семантикой выражения, лежащего в ее основе.

Работа продолжает ряд исследований (в том числе сопоставительных) семантики идиом, проводимых нами ранее (ср. работы (3), (4), (5)), а описания представленных ниже идиом являются частью проекта Отдела экспериментальной лексикографии Института русского языка РАН им. В. В. Виноградова по созданию серии словарей современной русской идиоматики и приводятся в формате, используемом в Фразеологическом объяснительнооом словаре русского языка (см. подробнее об устройстве Словаря в (2)).

1. Для начала сравним, как данные идиомы представлены в существующих современных фразеологических словарях. В словаре (Телия 2006) все вышеперечисленные идиомы отсутствуют; в других словарях (см. список рассматриваемых фразеологических словарей в конце работы) приводится лишь часть из них:

	альфа и омега	*от альфы до омеги*	*от а до я*	*от аза до ижицы*	*от а до ижицы*	*от а до зет*
Лубенская 1997	+	+	+	+	–	+
Мокиенко 1997	+	–	–	–	–	–
Яранцев 2001	+	–	–	–	–	–
ФСРЯ 2002	+	+	+	–	+	–
Степанова 2003	+	+	+	–	–	+
Телия 2006	–	–	–	–	–	–

2. Рассмотрим далее предлагаемые толкования идиомы *альфа и омега*. Во всех цитируемых словарях у идиомы выделяется два значения (в (Мокиенко 1997) они объединены в одно):

1) Начало и конец

2) Сущность, основа, самое главное

При этом на первое значение (начало и конец) примеры либо не приводятся (ср. Лубенская 2007), либо эти примеры устаревшие[1], либо они не соответствуют заявленному значению. Ср. следующие примеры из словаря Яранцев 2001:

(91) Первым и последним словом, *альфой и омегой* всей его жизни было, как у всех поэтов, — его собственное я. И. С. Тургенев. Рецензия на «Фауст» Гете

[1] Современными примерами мы считаем примеры из литературы, публицистики или речевого общения не ранее 60-х годов 20-го века.

(92) Для многих ученых диссертация стала *альфой и омегой* их научной карьеры, средоточием их помыслов на много лет вперед, иногда на всю активную жизнь. И. Бестужев-Лада, Середняк в науке

(93) — Вам когда-нибудь доводилось быть пайщиком жилкооп-товарищества? Нет? Значит, вы не знаете жизни. Современный жилищный кооператив это, если хотите знать, *альфа и омега*, начало всех начал в текущем быту интеллигента средней руки. В. Максимов. Карантин

Из приведенных примеров только (2) может быть проинтерпретирован как употребление идиомы в первом значении, ср.: *Для многих ученых диссертация стала началом и концом их научной карьеры.* Впрочем, этот же пример может быть проинтерпретирован и как употребление идиомы во втором значении, ср.: *Для многих ученых диссертация стала самым главным в их научной карьере, средоточием их помыслов на много лет вперед, иногда на всю активную жизнь.*

Среди многочисленных рассмотренных нами примеров употребления этой идиомы[2] был обнаружен только один, в котором можно усмотреть реализацию первого значения:

(94) Он уже быстро срывал бумажные листы один за другим со свёртка и наконец бросил на стол чёрный, обугленный предмет — не то камень, не то хлебную корку, неправильно-сферической формы. Затем снял газету с «красной леди» и положил на неё и эту обуглившуюся, покоробленную временем кость. Так они лежали вместе — большой розовый череп и чёрный круглый костяной обломок. — Ну вот и гейдельбержец тут, — сказал он ублаготворенно, — вся компания, значит, *альфа и омега*. Низшие и высшие. Все тут, у меня на столе. Теперь пусть разбираются, кто от кого. Ю. О. Домбровский. Обезьяна приходит за своим черепом

Однако, по нашему мнению, в данном случае имеет место нестандартное употребление: был реализован потенциал членимости внутренней формы идиомы — *альфа* и *омега* в данном случае указывают на 2 разных субъекта, в то время как при стандартном употреблении *альфа и омега* согласуется с одним субъектом [ср. примеры (1–3)].

Как следствие из вышесказанного, далее мы будем считать, что идиома *альфа и омега* не используется в современном русском языке в значении ‘начало и конец’ (если интерпретировать эти слова в их прямом, а не переносном значении, что является одним из основных требований к словам, используемым в толковании). Очевидно, вводя первое значение идиомы, авторы соотвуетствующих словарей стремились передать метафору, лежащую в основе идиомы и впервые встречающуюся в новозаветном библейском тексте, к которому и восходит современная идиома:

(95) Я есмь Альфа и Омега, начало и конец, говорит Господь, Который есть, и был, и грядет, Вседержитель. Откровение святого Иоанна Богослова, 1.8

Альфа и омега, первая и последняя буквы греческого алфавита, в данном случае переосмысляются как начало и конец не алфавита, но всего сущего, а также континуум между ними — т. е. все, что есть вообще. ср. *Я — все* (т. е. все существенное, все, что важно)[3]. Неслучайно многие примеры (в основном из литературы 19в.) приводят через запятую выражение «начало и конец», очевидно, понимаемое метафорически, ср. примеры, приводимые в (Мокиенко 1997) и (ФСРЯ 2002):

(96) У нее в мире никого нет, кроме него, он для нее все — закон, родство, природа, начало и конец, *альфа и омега* ее бытия, все, все. И. Лажечников. Ледяной дом

(97) Вот вам *альфа и омега*, начало и конец всех дел, имевших цель удовлетворитьте или другие общественные потребности. Г. Успенский. Бог грехам терпит

2.1. Остановимся подробнее на значении идомы ‘сущность, основа, самое главное’. В дальнейшем мы будем говорить о нем как о единственном значении данной идиомы в русском языке[4]. Вот как оно протолковано в нашем Словаре (2) на основе анализа большого корпуса примеров употребления:

АЛЬФА И ОМЕГА (чего-л.) *книжн.* самая существенная часть чего-л., без которой его не было бы как такового, *подобно тому, как не существовал бы алфавит без его первой и последней букв* ❖ основа, суть, сущность.

---

[2] Поиск примеров производился в Национальном корпусе русского языка, Базе данных по русской фразеологии, а также в русскоязычном сегменте Интернета.

[3] Ср. также комментарий в (Степанова 2003): Выражение пришло в русский из апокалипсиса, где Бог говорит: «Аз есть альфа и омега», т. е. я заключаю в себе все, что может быть названо на буквы греческого алфавита с первой до последней, т. е. я — начало и конец всего на свете.

[4] В богословских текстах и на иконах *α и ω, Альфа и Омега* используются как одно из имен Бога (ср. Спаситель, Отец, Вседержитель и т. п.). В английском языке также выделяется такое значение, ср. The Oxford English Dictionary: Today people still sometimes use the capitalized form, Alpha and Omega, to denote the Divine Being. Это значение могло бы быть представлено с пометой *спец.*, которая на настоящий момент не предусмотрена в нашем Словаре, либо в рамках отдельного комментария в соответствующей словарной статье.

В толковании элементы значения, унаследованные от внутренней формы идиомы, выделены курсивом; затем приведены однословные квазисинонимы идиомы[5]. Ср. типичные примеры употребления этой идиомы:

(98) Российское — советское — русское государство всегда двигалось вперед за счет ограбления своего собственного народа. Это — суть, квинтэссенция, *альфа и омега* даже и реформаторской деятельности в России, не говоря уже о нереформаторской. Независимая газета, 1996

(99) *Альфа и омега* марксизма — постоянная борьба, та самая перманентная революция, которой всегда следовали большевики с тех пор, как они вышли на баррикады первой русской революции. А. Яковлев. Омут памяти

Таким образом, относительно первоначальной метафоры наблюдается сужение значения: от значения ‘все важное, существенное’ к значению ‘самое важное, существенное в некоторой области’. Во внутренней форме идиомы осталась отсылка к первой и последней буквам алфавита как к способу номинации смысла ‘самое важное’. При этом в современном употреблении субъектная валентность не может заполняться обозначением человека. Единичные контексты такого рода встречаются в литературе 19в. (ср. контекст (6) выше), а также в текстах религиозного содержания или обсуждающих религиозные темы, однако в целом для русского языка такое употребление нетипично, ср.:

(100) Тот, кто приходит к отцу Александру, кто сейчас с ним, для него — *альфа и омега*, все на свете. У батюшки происходит некий акт глубокой самоотдачи. С каждым приходящим возникают особые, непохожие на другие отношения. Человек чувствует себя любимым, интересным, важным, более того — единственным. З. Масленикова. Жизнь отца Александра Меня

В современном употреблении идиомы субъектная валентность и дополнение в родительном падеже стандартно заполняются элементами непредметной лексики (как правило, абстрактными понятиями), ср. необычность в русском языке таких выражений, как **фрукты — альфа и омега любой диеты*; **стулья — альфа и омега мебельной промышленности*[6]. В прототипическом случае речь идет о некоторой понятийной области, выраженой дополнением, в которой выделяется меньшая понятийная область, ‘самое важное’, которая описывается субъектом.

В качестве исключения из правила заполнения субъектной валентности непредметной лексикой можно привести игровой иронический контекст употребления идиомы, где эффект комичности достигается в том числе за счет нарушения этого правила:

(101) Граждане! Уважайте пружинный матрац в голубых цветочках! Это — семейный очаг, *альфа и омега меблировки*, общее и целое домашнего уюта, любовная база, отец примуса! Как сладко спать под демократический звон его пружин! Какие сладкие сны видит человек, засыпающий на его голубой дерюге! Каким уважением пользуется каждый матрацевладелец! И. Ильф, Е. Петров. Двенадцать стульев

3. В то время как выражение *альфа и омега* обозначает некоторую сущность, а соответствующая идиома выполняет в предложении роль предиката (что-л. является *альфой и омегой* чего-л.), выражения вида *от альфы до омеги* (*от А до Я* и т. п.) указывают на протяженность, а соответствующие идиомы в предложении являются сирконстантом и обозначают свойство действия (некоторое действие выполнено *от альфы до омеги*):

(102) Меня порой спрашивают, какое из виденных мною в цирке представлений можно назвать спектаклем в полном смысле этого слова, и я всегда в подобных случаях вспоминаю спектакль Андрея Николаева, где всё было продумано *«от А до Я»*. И. Э. Кио. Иллюзии без иллюзий

Словари (ФСРЯ 2002) и (Степанова 2003), в которые вошли идиомы этого ряда, описывают их значение одинаково: от самого начала до конца. В словаре (Лубенская 1997) дается определение: (прочитать, знать *что* и т. п.) from the very beginning to the very end; (to know sth.) thoroughly. С нашей точки зрения, как и для идиомы *альфа и омега*, выражение «от самого начала до конца» описывает не собственно значение идиомы, а переданный метафорически способ номинации смысла ‘полностью’, ср. далее, как это будет отражено в нашем толковании.

Заметим, что в словаре (Лубенская 1997) сочетаемость идиом ряда *от альфы до омеги* ограничена глаголами *читать, знать* и т. п., сочетающимися с идей алфавита во внутренней форме идиомы: читать и знать можно некоторый текст, информацию, которая можно быть исчислена алфавитным списком. Вероятно, какое-то время назад эти идиомы употреблялись только с подобными глаголами, ср. пример из (ФСРЯ 2002):

---

[5] Подробнее о формате толкования см. (2), О пользовании словарем.

[6] Примечательно, что в английском и немецком языках это не так, см. подробнее о различиях в значении и употреблении идиом *альфа и омега, alpha and omega, alpha und omega* в русском, английском и немецком языках в работе (5).

(103) Впрочем, желаю ему доброго здоровья и не завидую в надежде читать записки *от альфы до омеги* Н. М. Карамзин. Письмо И. И. Дмитриеву

Однако судя по большому корпусу современных контекстов, в настоящее время эти идиомы употребляются и с действиями и процессами, не имеющими непосредственного отношения к информации, ср.:

(104) Процесс интеграции таких [авиационных] предприятий широко практикуется в мире. Подобная же практика существует и на Западе, где нет крупных фирм, занимающихся отдельно разработкой или производством. Это единый технологический цикл, где весь процесс — *от «а» до «я»*, находится в одних руках. Известия, 2003

(105) В последнее время появилось много архитектурных студий и мастерских, предлагающих так называемый «комплексный архитектурный подход» к строительству загородного дома. То есть создание гармоничного жилища *от А до Я*: начиная с разработки места застройки, заканчивая покупкой настенных часов и прочими интерьерными нюансами. Мир & Дом. Residence, 2004

Однако какие-то ограничения на виды деятельности, могущие быть описанным такими идиомами, все же остаются. Так, в примерах (14), (15) и им подобных речь идет о какой-то сложносоставной деятельности, этапы которой можно представить списком от А до Я. Ср. сомнительность употреблений **перекопать весь участок от А до Я, *помыть окно от А до Я*. Исходя из ограничений на употребление идиомы, представляется целесообразным в данном случае выделить 2 значения в зависимости от сочетаемости с глаголами соответствующих классов:

ОТ А ДО Я 1. (сделать что-л.) сделать что-л. полностью, *как бы выполнив все части процесса от первой до последней по алфавитному списку* 🗐 Для того чтобы Америка доверилась России, наши технологии должны стать полностью прозрачными. Это немыслимо — показать кому-либо, как *от «а» до «я»* делается и разбирается наша атомная бомба. Последний наш прикуп перед всем миром… Известия, 2001

ОТ А ДО Я 2. читать, знать, помнить… (что-л.) читать, знать, помнить и т. п. какую-либо информацию полностью, *как бы от первого до последнего элемента по алфавитному списку* 🗐 Он знает абсолютно все про все, особенно про американское искусство. Обо всем у него есть свое мнение, большей частью американское искусство он ненавидит, особенно его массовую часть, однако влюблен в американский мюзикл и знает его *от «А» до «Я»*. А. Журбин. Как это делалось в Америке. Автобиографические заметки

Во внутренней форме идиомы сохранился способ номинации значения ‘полностью’, учтенный в обоих толкованиях.

Так же как и авторы перечисленных словарей, мы считаем идиомы *от альфы до омеги, от а до я, от аза до ижицы, от а до зет* синонимичными, имеющими лишь стилистические различия. Так, идиома *от альфы до омеги* крайне малоупотребительна в современном языке и является устаревшей, как и идиома *от аза до ижицы*. Эти идиомы в современной литературе встречаются в стилизованных контекстах:

(106) Каждую мысль он стремился воспринять и сформулировать так, чтобы в ней выразился и зазвучал весь человек, тем самым в свернутом виде все его мировоззрение *от альфы до омеги*. М. М. Бахтин. Проблемы поэтики Достоевского

(107) Лазарева суббота. Закончилась Всенощная. Мы, учащиеся молодежной группы, собрались в здании Воскресной школы и по-братски разделили баночку икры, дабы "от аза до ижицы" исполнить устав церкви. ХудТексты Инт.

Идиома *от а до зет* не является устаревшей, однако встречается достаточно редко, часто в переводных текстах, или в текстах, где важен контекст англоязычной культуры (или другой культуры с латинской письменностью):

(108) ДиректорЦРУлюбил, когдаемурастолковывали вопрос от А до Зет, как будто он сам не знал никаких подробностей. Чейз Дж. Х. Предоставьте это мне.

Отметим, что в Интернете помимо рассмотренных синонимичных идиом встречаются также многочисленные игровые модификации, такие как *от А до Ю, от А до Ы, от А до Щ, от Б до Ю* и т. п., ср.: Организация праздников от «Б» (банкетов) до «Ю» (юбилеев); название книги «Коррупция от А до Ю», автор Г. Явлинский. Разумеется, такие употребления мы считаем нестандартными.

3.1. Помимо упомянутых двух значений идиом вида *от альфы до омеги* мы выделяем третье значение; в этом значении такие идиомы встречаются в названиях книг и заголовках, ср.: *Справочник необходимых знаний от Альфы до Омеги; Америка от А до Я.; Отопление от А до Я.; Менеджмент и финансы от А до Я*. Ср. предлагаемое нами толкование для этого значения идиомы *от А до Я*:

ОТ А ДО Я 3. (что-л.) полная информация о какой-л. тематической области, *как бы от первого до последнего элемента по алфавитному списку*

В части толкования, отражающей внутреннюю форму идиомы, здесь также сохраняется способ номинации значения ‘полностью’.

Заметим, что и в данном случае речь может идти только об информации, — которая представима в упорядоченном виде. Так, выражение *горный козел от А до Я* может означать ‘все о горном козле’ и не может ‘весь горный козел’, ср., *горный козел от рогов до копыт*. Таким образом, мы снова наблюдаем, как внутренняя форма идиомы задает ограничения на ее сочетаемость.

В этом значении идиома *от А до Я* также наиболее частотна, реже встречается *от альфы до омеги*, встречаются единичные примеры употребления идиом *от аза до ижицы* и *от а до зет* в этом значении[7].

4. Итак, в работе были рассмотрены идиомы *альфа и омега, от альфы до омеги, от а до я, от аза до ижицы, от а до зет*. Было предложено четыре толкования, с нашей точки зрения полно описывающих все имеющиеся в современном русском языке значения этих идиом:

АЛЬФА И ОМЕГА (чего-л.) *книжн.*
ОТ А ДО Я 1. (сделать что-л.)
ОТ А ДО Я 2. читать, знать, помнить… (что-л.)
ОТ А ДО Я 3. (что-л.)

Аналогично идиоме **ОТ А ДО Я** описываются идиомы **ОТ АЛЬФЫ ДО ОМЕГИ** *устар.*, **ОТ АЗА ДО ИЖИЦЫ** *устар.*, **ОТ А ДО ЗЕТ.**

Было показано, как современное значение этих идиом связано с первоначальной метафорой, почему многие словари приводят несуществующие значения этих идиом, какие элементы внутренней формы включены в толкование и как они влияют на сочетаемость идиомы с теми или иными классами глаголов. Наше исследование служит еще одним аргументом в пользу того, что внутренняя форма идиомы — не что-то мертвое и застывшее; она ощущается носителем языка и позволяет более полно описать значение идиомы, а также объяснить ограничения на ее употребление.

[7] Данные о частотности употребления идиом приводятся соответственно количеству найденных в русскоязычном сегменте Интернета контекстов.

Литература

1. *Баранов А. Н., Добровольский Д. О., Киселева К. Л. Козеренко А. Д.* при участии М. М. Вознесенской и М. М. Коробовой, под редакцией А. Н. Баранова и Д. О. Добровольского. Словарь-тезаурус современной русской идиоматики, М., 2007
2. *Баранов А. Н., Вознесенская М. М., Добровольский Д. О., Киселева К. Л., Козеренко А. Д.* Фразеологический объяснительный словарь русского языка, М., 2009
3. *Козеренко А. Д.* Внутренняя форма идиом в сопоставительном аспекте // Труды Международного семинара Диалог’2003 по компьютерной лингвистике и ее приложениям. Протвино, 2003
4. *Козеренко А. Д.* Скобки в русских идиомах // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2009» Вып. 8 (15). — М.: РГГУ, 2009.
5. *Козеренко А. Д.* Альфа и омега: идиомы с одинаковой внутренней формой в русском, английском немецком языках // Oslo Studies in Language 2 / 2010 (в печати)

Словари и корпуса текстов

1. *Лубенская 1997* — Лубенская С. И. Русско-английский фразеологический словарь. М., 1997
2. *Мокиенко 1997* — Мелерович А. М., Мокиенко В. М. Фразеологизмы в русской речи. Словарь. М., 1997
3. *Яранцев 2001* — Яранцев Р. И. Русская фразеология. Словарь-справочник: Ок. 1500 фразеологизмов. М., 2001
4. *ФСРЯ 2002* — Фразеологический словарь русского языка // Составители: Л. А. Войкова, В. П. Жуков, А. И. Молотков, А. И. Федоров / Под ред. А. И. Молоткова, издание третье, стереотипное. — М, 2002
5. *Степанова 2003* — Фразеологический словарь / Составитель Степанова М. И. СПб.: Виктория плюс, 2003.
6. *Телия 2006* — Большой фразеологический словарь русского языка. Значение. Употребление. Культурологический комментарий / Отв. Ред. В. Н. Телия. М., 2006
7. *The Oxford English Dictionary* — The Oxford English Dictionary, Second Edition, Oxford, 2005
8. *Национальный* корпус русского языка
9. *База данных* по русской фразеологии (ИРЯ им. В. В. Виноградова РАН, Отдел экспериментальной лексикографии)

Эволюция лингво-семантических представлений в интеллектуальных системах на основе расширенных семантических сетей

Evolution of Linguistic Semantic Presentations in the Intelligent Systems Based on the Extended Semantic Networks

Козеренко Е. Б. (kozerenko@mail.ru),
Кузнецов И. П. (igor-kuz@mtu-net.ru)
Институт проблем информатики РАН, Москва

В работе рассматриваются вопросы проектирования и развития семантико-синтаксических и лексико-семантических представлений в лингвистических процессорах ряда систем, основанных на аппарате расширенных семантических сетей (РСС). Системы этого класса создаются для извлечения знаний из текстов на естественных языках, отображения извлеченных сущностей и связей в структуры базы знаний и использования знаний для поддержки экспертных аналитических решений в различных сферах приложения. В фокусе внимания находятся инженерно-лингвистические представления, позволяющие построить целостную работающую лингвистическую модель, которая модифицируется в зависимости от конкретной задачи: от «тяжелой» формы на основе детальных глубинных представлений до фокусных редуцированных оболочек, настроенных на узкую предметную область и ограниченный язык общения.

1. Введение

Данная работа посвящена вопросам создания инженерно-лингвистических моделей естественного языка для построения лингвистических процессоров различных классов информационных систем и описанию опыта создания лингвистических представлений в системах, относящихся к области исследований искусственного интеллекта. В центре нашего внимания находятся интеллектуальные системы, разработанные на основе аппарата *расширенных семантических сетей* (РСС) [1–3, 18–19]. Мы будем их называть *РСС-системы*. Эти системы создавались коллективом разработчиков, включая авторов данной статьи в Институте проблем информатики РАН на протяжении целого ряда лет в рамках исследовательских проектов и прикладных систем, ориентированных на конкретные предметные области заказчиков. Мы выделяем 4 поколения РСС-систем. Лингво-семантические представления, заложенные в основу систем этого класса прошли определенный эволюционный путь.

Интеллектуальные РСС-системы содержат развитые *базы знаний*, при этом знания представлены в виде записей на языке расширенных семантических сетей, называемых *РСС-структурами*. Лингвистические знания, таким образом, являются частным случаем «знаний» и также представлены в виде записей на языке расширенных семантических сетей. Основным конструктивным элементом РСС является именованный N-местный предикат, называемый «*фрагментом*». Все множество языковых объектов задается в виде системы предикатно-актантных структур, при этом поддерживаются механизмы представления вложенных структур, что дает очень мощные изобразительные возможности для описания объектов различных языковых уровней. Очень важным фактором является однородность и единообразие лингвистических представлений.

В процессе анализа и синтеза предложений естественного языка используется формально-грамматический аппарат, сходный с грамматиками зависимостей. При этом подходе опорными элементами являются слова и конструкции, выполняющие

роль предикатов в предложении, и результатом анализа предложения должен стать один предикат, соответствующий сказуемому рассматриваемого предложения (т. е. основному глаголу в личной форме или другому основному предикатному выражению). Таким образом, в процессе анализа, в первую очередь, происходит выявление «слов-действий» и «слов-отношений», т. е. глаголов и других слов, имеющих синтактико-семантические валентности. Примером «слов-отношений» могут служить, например, слова «отец», «друг», и т. п., то есть в данном случае «отношения» — это слова, которые задают сильные четко выраженные синтактико-семантические ожидания.

Семантический анализ в инженерно-лингвистическом понимании — это процесс перевода естественно-языковых выражений во «внутренние» структуры базы знаний (БЗ), в нашем случае этими «внутренними» структурами являются записи на языке РСС. Таким образом, структуры БЗ — это код смысла в интеллектуальных информационных системах подобного рода.

В работе рассматриваются инженерно-лингвистические решения в системах с «полным» лингвистическим анализом — это системы 1-го и 2-го поколений: ДИЕС1, ДИЕС2, Логос-Д [2–3] и системах с «фактографическим» подходом — интеллектуальных системах поддержки аналитических решений (ИСПАР) [18–19], где целью анализа является выделение сущностей и связей из текстов — это системы 3-го и 4-го поколений.

2. Концептуально-лингвистическое моделирование в РСС-системах

2.1. Основные аспекты семантического моделирования

Концептуально-лингвистическое моделирование (КЛМ) — это процесс построения естественно-языковой модели предметной области (ПО) (Рис.1), синтезирующий в себе подходы концептуального и лингвистического моделирования [1–3]. Построение концептуально-лингвистической модели некоторой предметной области подразделяется на следующие этапы:

- построение собственно концептуальной модели, т. е. вычленение базовых понятий, организация их в родо-видовые деревья и определение связей между ними;
- разработка идеографического словаря предметной области, т. е. лексическое наполнение концептуальной модели;

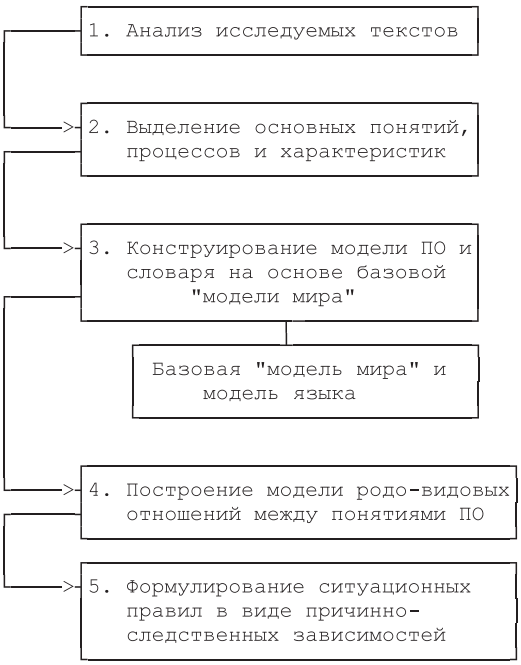


Рис. 1. Процесс концептуально-лингвистического моделирования. — ввод базовых правил, описывающих на естественном языке «модель мира», релевантную для данной ПО

Методика концептуально-лингвистического моделирования на основе аппарата РСС базируется на следующих принципах:

- модель должна быть «открытой», то есть поддерживать эффективный механизм расширения и обновления информации;
- модель представления «смысла» должна учитывать факты экстралингвистической реальности, которые в виде правил и отношений составляют некоторую базовую «модель мира», достраиваемую конкретными моделями предметных областей;
- модель должна быть практичной, то есть не перегруженной детальными описаниями связей и отношений между понятиями, чтобы обеспечить возможность ее реализации, но в то же время, отражать всю релевантную для конкретной задачи информацию.

Реалистичный подход к постановке задачи диктует необходимость ограничения моделируемого подмножества естественного языка. Суть ограничений сводится к следующему:

- во-первых, анализируемые текстовые материалы содержат экспертные знания из конкретных предметных областей (в разработанных авторами системах это были такие предметные области как диагностика брака при изготовлении микросхем, социальное прогнозирование, криминалистика, и другие);

- во-вторых, в целях максимально возможного устранения неоднозначности, словарь строится по модульному принципу: есть некоторая наиболее общая часть (1–2 уровня), которая достраивается специальными словарями для каждой отдельной предметной области.

Предлагаемая модель лексической семантики основана на принципе «ядерного» значения, реализуемого в контексте данной предметной области, с последующим индуктивным наращиванием других значений (если они актуализируются в рассматриваемых контекстах). Также используется таксономия которая реализуется в виде иерархических деревьев классов слов.

Общая «модель мира» системы служит основой для моделей ПО. Элементами этой модели являются классы слов, которые подразделяются на

- понятия / имена,
- отношения,
- действия,
- свойства,
- характеристики действий,
- временные и пространственные характеристики.

Самым общим понятием является *концепт*, или *универсальный класс*, который подразделяется на объект, ситуацию, процесс и др.

Слова, относящиеся к классам действий и отношений, представлены как семантико-синтаксические фреймы, задающие предикатно-актантные структуры (модель управления). Однако, в описываемом подходе (назовем его РСС-подход) существенно расширена область значений актантов. Суть расширения состоит, во-первых, в том, что в роли актантов могут выступать не только простые объекты, соответствующие отдельным словам, но и структурные объекты, представляющие словосочетания и фразы, а во-вторых, в том, что понятие «падежа» включает в себя не только семантические, но и синтаксические признаки.

Подход, основанный на РСС, позволяет отражать произвольный уровень вложенности структур за счет пропозициональных вершин семантической сети, что обеспечивает представление сложных синтаксических конструкций фраз ЕЯ, а также позволяет отразить структурный характер лексической семантики, которая в предлагаемой модели имеет иерархически-сетевую структуру. Лингвистические знания представлены в системном словаре и декларативных модулях лингвистического процессора. В РСС-системах также реализована функция динамически формируемого семантического словаря, который на основе исходной лингвистической информации достраивается системой автоматически в процессе обработки конкретных текстов. На Рис. 2 представлено такое «внутреннее» описание глагола в семантическом словаре. Этот словарь автоматически генерируется РСС-системами ДИЕС2, ЛОГОС-Д, ИКС в процессе обработки естественно-языковых текстов.

```
{(ВЫРАБАТЫВА895__)(DICSEM)
 COORD(PROGNOZ1,RUS,ВЫРАБАТЫВА895__,S50_31_51_20,%)
          SUB(UNIV,0+) SUB(UNIV,1+) SUB(UNIV,2+)
 ВЫРАБАТЫВ(0-,1-,2-/3+) INFI(3-) ПРИДЕТСЯ(3-)
                                ПРИДЕТСЯ(3-/4+) FUT1(4-) SUB(СРЕД,5+)
```

Рис. 2. Пример записи представления глагола «вырабатывать» в семантическом словаре

2.2. Аппарат РСС — основа концептуально-лингвистического моделирования

Дадим краткое описание аппарата расширенных семантических сетей и дадим обоснование выбора именно этого метода представления для моделирования естественного языка. Классическое понятие семантической сети сводится к следующему: задаются некоторые вершины, соответствующие объектам. Вершины связываются дугами, которые помечаются именами отношений. Однако с помощью подобных сетей оказывается трудно представлять сложные виды информации, например, когда объекты, связанные отношениями, образуют агрегаты, и когда отношения связываются между собой отношениями и др. Поэтому в сети вводятся вершины, соответствующие именам отношений, а также специальный композиционный элемент, называемый вершиной связи. Вершина связи как бы «разрывает» дугу и подсоединяется одним ребром к вершине-отношению, а другими ребрами — к вершинам-объектам. РСС является развитием такого сорта сетей в направлении повышения изобразительных возможностей при сохранении свойства однородности.

Основой РСС является множество вершин (V), из которых составляются элементарные фрагменты (ЭФ) следующего вида:

$$V_0(V_1, V_2, \ldots, V_k/V_{k+1}),$$

где $V_0, V_1, V_2, \ldots, V_k, V_{k+1}$ V, $k > 0$.

Такой фрагмент представляет k-местное отношение. Позиции вершин в элементарных фрагментах (ЭФ) определяют их роли. Вершина V_0 ставится в соответствие имени отношения, вершины V_1, V_2, ..., V_k — объектам, участвующим в отношении, а вершина V_{k+1}, отделенная косой линией (/), — всей совокупности упомянутых объектов с учетом их отношения. В дальнейшем будем V_{k+1} называть С-вершиной ЭФ. Множество ЭФ образуют расширенную семантическую сеть (РСС). С помощью РСС представляются наборы отношений, различные ситуации, сценарии. Сильной стороной РСС-подхода является возможность однородного представления как предметной (концептуальной), так и лингвистической информации, что обеспечивает эффективную обработку знаний и поддержание непротиворечивости базы знаний.

Посредством РСС в базе знаний представлены лингвистические (ЛЗ) и предметные знания (ПЗ). Обработка этих знаний осуществляется продукциями языка ДЕКЛ, на котором реализованы следующие шесть блоков: морфологического анализа (МА), семантического анализа слов (САС), синтактико-семантического анализа форм (ССА), прагматических функций (ПФ), организации системной активности (БА) и обратный лингвистический процессор (ОЛП). С помощью продукций осуществляется последовательное преобразование сети — РСС. При этом проходятся фазы, соответствующие уровню понимания входного текста. Рассмотрим их.

1. На первом шаге анализа происходит построение пространственной структуры предложения с морфологической информацией для каждого слова. Каждый член предложения представляется вершиной семантической сети. Вместо слова — генерируется код (если слово многозначно, т. е. принадлежит к нескольким классам, — то более одного кода). Основой кода служит корень слова. На этом этапе предложение представляется в виде набора фрагментов типа LRR (специальные метки результатов 1-го этапа анализа), объединяемых в целостную структуру посредством вершины связи. Результат 1-го этапа постоянно обращается к словарю: «Что значит данное слово?»
2. На втором этапе каждой вершине сопоставляется семантический класс и присваивается новый код. За словами (т. е. конкретными вершинами РСС) система видит объекты, действия, свойства — то есть, строит классификации. Производится семантико-синтаксический анализ без выявления глагольных словоформ, при этом предложение представляется в виде совокупности фрагментов типа SEM и SEMD (специальные метки результатов 2-го этапа анализа) (Рис. 3).
3. На третьем этапе происходит частичное «сворачивание» синтаксических структур в более компактные (например, свойство объекта и сам объект) с присваиванием нового кода, и строится фрагмент для объекта, обладающего эти свойством.
4. На четвертом этапе выявляются отношения и действия и производится анализ непосредственного контекста на соответствие заданным семантическим падежам. Система смотрит, подходят ли объекты (концепты, понятия) на аргументные места данного действия или отношения. При этом отглагольные существительные («делатель» — т. е. агент действия, или «деланиe» — процесс, анализируются как слова с двойной природой — вначале как действия, а затем как объекты). Результатом этого этапа является целостная семантическая структура предложения, которая представляется фрагментом типа SEMSTR (метка результата 4-го этапа анализа) (Рис. 4).
5. На пятом этапе происходит анализ прагматики: установление кореферентциальных отношений, частичное восстановление эллиптических конструкций, система производит дальнейшие действия с построенными фрагментами.

Рис. 3. Семантико-синтаксический анализ без выявления глагольных словоформ

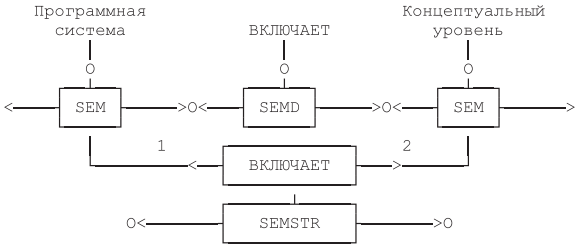


Рис. 4. Целостная семантическая структура предложения

ДИЕС допускает ввод полисемичных форм глаголов. Для этого следует воспользоваться формальной записью лингвистических знаний. Например, можно ввести запись: ВЗЯЛ/-ДЕЙСТВИЕ, КОГО-ЧЕЛОВЕКА ЗА ЧТО-ПРЕСТУПЛЕНИЕ.

Тогда ДИЕС будет понимать предложения типа ИВАНА ВЗЯЛИ ЗА КРАЖУ и другие предложения такого типа. Но ДИЕС будет отличать это действие от других значений глагола ВЗЯТЬ, например, ВЗЯТЬ КНИГУ. Итак, в системах, основанных на РСС, все функции реализованы на единой основе — в рамках языков РСС и ДЕКЛ, которые были разработаны с ориентацией на задачи обработки естественного языка.

3. Представление семантики глаголов, глубинные и поверхностные структуры

В процессе анализа выявляются семантические вершины предложения — происходит выявление «слов-действий», т. е. глаголов, и «слов-отношений». Что же является конструктивной основой задания семантических представлений предикатных слов и выражений? Как убедительно показано в работе Ю.Д. Апресяна «Экспериментальные исследования семантики русского глагола» [4], семантика глагола определяется его дистрибутивно-трансформационными свойствами. Поэтому смысл предикатных выражений должен кодироваться с учетом их дистрибутивных и трансформационных признаков.

Выдвинутая рядом лингвистов гипотеза (Хомский, Филлмор) [5–8] о том, что все предложения имеют глубинные и поверхностные структуры, явилась очень продуктивным источником проектных решений при создании первых РСС-систем и развивалась в дальнейшем. В теоретико-лингвистическом понимании глубинная структура — это абстракция, содержащая все элементы, необходимые для образования поверхностных структур предложений со сходной семантикой. В инженерно-лингвистическом понимании глубинная структура — это запись на языке БЗ, например, на РСС, которая может быть представлена в «поверхностном» виде на одном из естественных языков в результате конечного числа определенных преобразований. Например, предложения

(109) The dog chases the cat.
(110) The cat is chased by the dog.

имеют истоком одну глубинную структуру:

хотя и отличаются своими поверхностными структурами. В каждом из них имеется агент (the dog), объект (the cat), и действие (chase). Согласно концепции *падежной грамматики* Филлмора [5], глубинная структура для обоих предложений инвариантна. Эту структуру можно представить в виде скобочной записи V(AGENT, OBJECT). В графическом виде глубинная структура предложения также может быть представлена диаграммой в виде дерева, где отражены инвариантные отношения зависимости между предикатной вершиной и актантами (Рис. 5), при этом в таком представлении явным образом разграничиваются *модальность* (MOD) и *пропозиция* (PROP).

В исходном виде [5] теория признавала шесть падежей: агентив, инструменталис, датив, объектив, локатив и фактитив. По мере развития теории [8] происходило увеличение числа падежей, однако «умножение» количества падежей утяжеляет первоначальную конфигурацию, поэтому при построении инженерных семантических представлений требуется некоторый «компромиссный» вариант, сочетающий в себе необходимую полноту, с одной стороны, и простоту и гибкость, с другой.

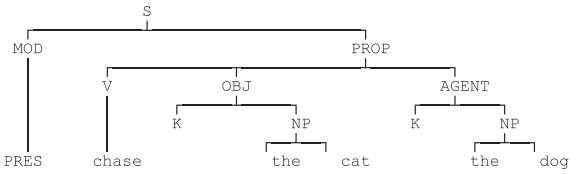


Рис. 5. Глубинная структура предложений.

4. Многоязычные системы

Одним из приоритетных направлений развития РСС-систем было обеспечение обработки текстов на нескольких языках, прежде всего, для русско-английской языковой пары. В системах 2-го поколения — ДИЕС2, ИКС, ЛОГОС-Д были реализованы лингвистические процессоры и словари для русского и английского языков, позволявшие обрабатывать тексты для ряда предметных областей, также поддерживались режим ввода лингвистических знаний лингвистом-аналитиком и автоматический режим самообучения системы по вводимым текстам. Проводились также эксперименты для итальянского и французского языков. При создании многоязычных систем мы обращались к европейским языкам. Очевидно, что европейские языки обладают большим количеством общих правил, чем любой из них с языками других групп. Но при этом все естественные языки обладают общей структурой на самом глубинном уровне. На этом уровне располагаются главные элементы естественного языка: Предложение, Модальность, Пропозиция.

Моделирование смысловых представлений — это процесс, развивающийся в направлении от поверхностных семантических структур — к глубинным. Поиск такого внутреннего представления смысла в условиях многоязычной ситуации является развитием методов концептуально-лингвистического моделирования на базе расширенных семантических сетей.

5. Интеллектуальные системы поддержки аналитических решений

РСС-системы 3-го и 4-го поколений направлены на извлечение знаний в виде *объектов,* или *сущностей,* и связей между ними из предметно-ориентированных текстов на русском и английском языках [18–19].

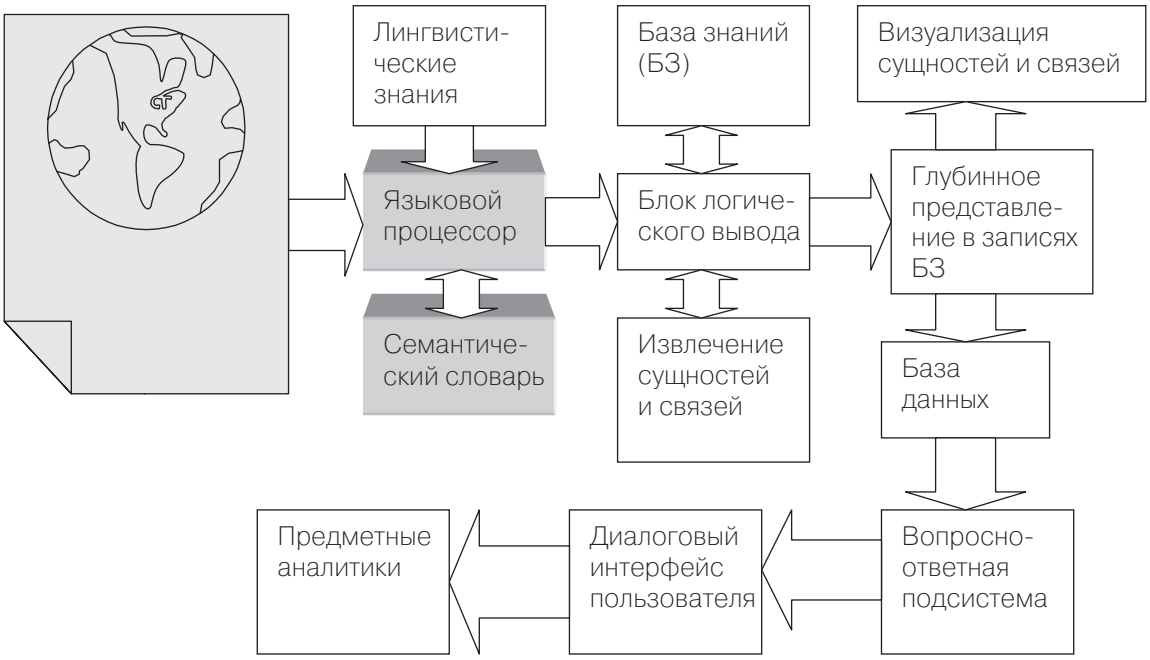


Рисунок 6. Обобщенное функциональное представление систем ИСПАР.

В настоящее время в мире активно ведутся работы по созданию систем извлечения фактов из текстов на естественных языках [13–16], создаются развитые тезаурусы и онтологии [17]. РСС-системы функционально шире, поскольку помимо возможностей извлечения фактов поддерживают механизмы логического анализа и экспертного вывода на основе извлеченных знаний. Системы такого рода являются интеллектуальными системами поддержки аналитических решений (ИСПАР). В целом это направление исследований требует дальнейшей проработки лексико-семантических представлений, создания предметно-ориентированных семантических словарей. Обобщенное функциональное представление систем ИСПАР дано на Рис. 6.

В рамках ИСПАР на основе расширенных семантических сетей (ИСПАР-РСС) были реализованы полномасштабные и пилотные проекты для ряда предметных областей: криминалистики, управления кадрами, мониторинга финансово-экономического кризиса, и других [18–19].

6. РСС-подход в лингвистических исследованиях

В настоящее время в рамках проектов, направленных на создание открытых лингвистических ресурсов [20] для научно-практических целей ведутся работы по выравниванию параллельных текстов научных статей, патентов и финансово-экономических текстов. В качестве одного из методов выравнивания используется РСС-подход, поскольку он позволяет отразить глубинно-семантический уровень языковых структур. На рисунке 7 представлен фрагмент первого этапа лингвистического анализа в многоязычных системах — для «идеальной» ситуации, когда структуры исходного текста и текста перевода практически совпадают, такая ситуация имеет место в меньшинстве случаев. Основные трудности возникают при наличии переводческих трансформаций в параллельных текстах. Особое внимание мы уделяем глагольно-именным трансформациям, например, явлению *номинализации*, поскольку она очень продуктивна для всех исследуемых нами языков.

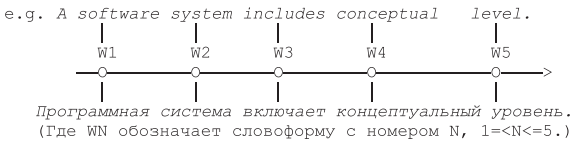


Рис. 7. Первый этап анализа параллельных текстов

Ключевой задачей при разработке методов сопоставления параллельных текстов является выявление и детальное описание тех языковых трансформаций, которые имеют место при переводе естественно-языковых конструкций с одного языка на другой [9], потому что далеко не всегда некоторое содержание передается структурно-подобными средствами в текстах на разных языках. Сравнительное исследование употребления различных частей речи в параллельных текстах на разных языках дает основу для выявления и описания языковых трансформаций, при этом центральной трансформацией является *номинализация*. Явление номинализации было исследова-

но в ряде работ отечественных и зарубежных лингвистов [9–12]. Ближе всего к нашему пониманию этого явления следующие определения номинализации: «конструкции… называются номинализованными — в том смысле, что их естественно рассматривать как результат номинализации конструкций с предикативным употреблением глаголов и прилагательных»; «номинализация — это синтаксический процесс, который соотносит предложения с именными группами». Выявление номинализованных конструкций в параллельных научных и патентных текстах на русском, английском, французском и немецком языках в научных и патентных текстах и сопоставительное описание глагольно-именных межъязыковых трансформаций — одна из центральных задач наших инженерно-лингвистических исследований.

7. Заключение

В данной работе представлен опыт создания и развития лингво-семантических представлений в интеллектуальных информационных системах, разработанных на основе аппарата расширенных семантических сетей (РСС). Аппарат РСС обеспечивает мощные изобразительные возможности для описания всех уровней естественного языка, включая уровень глубинно-семантических представлений, и межъязыковых соответствий. Конкретные лингвистические процессоры, которые были созданы на основе этого подхода, прошли определенный эволюционный путь и позволили выработать проектные решения для основных задач текущего этапа — извлечения и обработки содержательных знаний из текстов на естественных языках и сопоставления языковых структур в текстах на различных языках с учетом базовых трансформаций.

Проблема извлечения и обработки знаний открывает перспективы развития интеллектуальных направлений компьютерной лингвистики, поскольку ее основной акцент смещен в сторону глубинных представлений языка, в которых используются как грамматические (морфологические и синтаксические), так и семантические атрибуты для описания языковых объектов. Проводимые нами исследования параллельных текстов направлены также на рассмотрение этой проблемы [20]. Центральное место в наших лингвистических исследованиях занимает изучение и формализация процессов трансформации языковых структур, особенно все варианты глагольно-номинативных трансформаций, создание развитых дистрибутивно-трансформационных описаний предикатых структур для рассматриваемых языков.

Для задач извлечения знаний и создания систем ИСПАР дистрибутивно-трансформационные описания имеют также особое значение, поскольку таким образом задаются все возможные способы перевода языковых структур в предикатно-аргументные представления, которые затем используются в процедурах обработки знаний.

Литература

1. *Кузнецов И. П.* Семантические представления // Москва: «Наука», 1986. 290 с.
2. *Козеренко Е. Б.* Концептуально-лингвистическое моделирование в среде интеллектуального редактора знаний ИКС // «Проблемы проектирования и использования баз знаний». Ин-т кибернетики им. В.М. Глушкова, Киев, 1992. С. 73–79.
3. *Kozerenko E. B.* Multilingual Processors: a Unified Approach to Semantic and Syntactic Knowledge Presentation // Proceedings of the International Conference on Artificial Intelligence IC-AI'2001. H. R. Arabnia (ed.), Las Vegas, Nevada, USA, June 25–28, 2001. CSREA Press, 2001. P. 1277–1282.
4. *Апресян Ю. Д.* Экспериментальное исследование семантики русского глагола // Москва: Наука, 1967. 252 с.
5. *Филлмор Ч.* Дело о падеже // «Новое в зарубежной лингвистике». Вып. X. М.:Прогресс, 1968. С. 369–495.
6. *Хомский Н.* Аспекты теории синтаксиса // Москва: Изд-во МГУ, 1972.
7. *Хомский Н.* Язык и мышление// Москва: Изд-во МГУ, 1972.
8. *Fillmore C.* The case for case reopened // P. Cole & J.Sadok, Eds. Syntax and Semantics. New York: Academic Press. 1977. Vol. 8.
9. *Жолковский А. К., Мельчук И. А.* О семантическом синтезе // «Проблемы кибернетики», вып. 19. М, 1967.
10. *Падучева Е. В.* О семантике синтаксиса. Материалы к трансформационной грамматике русского языка. Изд. 2-е. // Москва: КомКнига, 2007. 296 с.
11. *Jacobs R. A.* and Rosenbaum P. S. English Transformational Grammar. // Blaisdell, 1968.
12. *Балли Ш.* Общая лингвистика и вопросы французского языка. Изд. 2-е, // Москва: УРСС, 2001.
13. *Cunningham H.* Automatic Information Extraction // Encyclopedia of Language and Linguistics, 2cnd ed. Elsevier, 2005.
14. *Han J. and Kamber M.* Data Mining: Concepts and Techniques // Morgan Kaufmann, 2006.
15. *FASTUS:* a Cascaded Finite-State Trasducerfor Extracting Information from Natural-Language Text. // AIC, SRI International. Menlo Park. California, 1996.
16. *Han J., Pei Y. Yin and Mao R.* Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach. // Data Mining and Knowledge Discovery, 8(1), 2004. P. 53–87.
17. *Добров Б. В., Лукашевич Н. В.* Онтологии для автоматической обработки текстов: Описание понятий и лексических значений // Компьютерная лингвистика и интеллектуальные технологии: Тр. междунар. конференции Диалог'06, Бекасово, 31 мая — 4 июня 2006 г., 2006. С. 138–142.
18. *Kuznetsov I. P., Efimov D. A., Kozerenko E. B.* Tools for Tuning the Semantix Processor to Application Areas // Proceedings of ICAI'09, Vol. I. WORLDCOMP'09, July 13–16, 2009, Las Vegas, Nevada, USA. — CRSEA Press, USA, 2009. P. 467–472.
19. *Kuznetsov I. P.,Kozerenko E. B.,Kuznetsov K. I.,Timonina N. O.* Intelligent System for Entities Extraction (ISEE) from Natural Language Texts // Proceedings of the International Workshop on Conceptual Structures for Extracting Natural Language Semantics — Sense'09, Uta Priss, Galia Angelova (Eds.), at the 17 International Conference on Conceptual Structures (ICCS'09), University Higher School of Economics, Moscow, Russia, 2009. P. 17–25.
20. *Kozerenko E. B.* INTERTEXT: A Multilingual Knowledge Base for Machine Translation // Proceedings of the International Conference on Machine Learning, Models, Technologies and Applications, June, 25–28, 2007, Las Vegas, USA. — Las Vegas: CSREA Press, 2007. P. 238–243.

Поиск ошибок в корпусе с помощью mte-разметки

Detecting errors in a corpus using mte-annotation

Копотев М. В. (mihail.kopotev@helsinki.fi)

Хельсинкский университет, Финляндия

В докладе описывается формат морфосинтаксического аннотирования MTE (MulText-East). На примере корпуса ХАНКО определяются возможности его применения для поиска ошибок морфологического аннотирования, выявляемых на основе анализа частоты совместной встречаемости граммем.

1. Описание формата

Формат MTE (MulText-East) — это многоязычный стандарт, разработанный для компактного и унифицированного представления морфосинтаксической информации для ряда европейских языков. Эта спецификация подготовлена группой из Любляны под руководством Т. Ержавеца для болгарского, чешского, эстонского, венгерского, румынского, словенского и английского языков (см. Erjavec [в печати]; Multext-East [электронный ресурс]). В 2007–2008 гг. стандарт был адаптирован к русскому языку (Sharoff et al., 2008). В настоящее время существует четвертая версия формата, продолжается работа по ее применению к другим славянским языкам (в том числе и другим восточнославянским).

Суть спецификации состоит в стандартном пословном морфосинтаксическом представлении в виде так называемых «тегсетов» (*tagsets*), или наборов морфологических показателей. Основная идея MTE заключается в том, что для каждой части речи представлена спецификация, устанавливающая набор атрибутов: **род, число, падеж** и т.д. с определенным списком значений (тегов): мужской, женский, средний. Место атрибута в тегсете фиксировано, его значение выражается буквой латинского алфавита. Первое место в тегсете всегда занимает показатель части речи, заданный заглавной буквой. Таким образом, морфосинтаксическое описание каждой текстоформы передается с помощью буквенного кода. Например,

Семеняка	Npfsny--	Семеняка
не	Qs	не
играла	V-is-sfa-p---	играть
в	Sps-	в
эти	Pd--paa--	этот
игры	Ncfpan--	игра
.	X	.

Для существительного (N) 1) собственного (p), 2) женского рода (f), 3) единственного числа (s), 4) в именительном падеже (n), 5) одушевленного (y) тегсет выглядит так:

Npfsny--

Если атрибуты по каким-то причинам не могут быть применены к определенной лексеме или текстоформе, место переменной заполняется дефисом. Например,

свято Afpns-s- святой

Прилагательное (A) качественно-относительное (f), в положительной степени (p), среднего рода (n), единственного числа (s) в краткой форме (s) не определено по атрибутам «падеж» (шестая позиция) и «сокращение» (восьмая позиция).

Спецификация MTE, представленная в работе (Sharoff et al. 2008), преследовала цели унифицировать формат для использования с разными языками и разными корпусами, что привело к уменьшению числа атрибутов. Следующая таблица дает представление о числе атрибутов для разных частей речи (см. табл. 1).

Два последних нуждаются в комментарии. Класс «Abbreviation» вызвал оживленную дискуссию при адаптации схему к русскому материалу, в результате которой было решено собрать в этот класс все аббревиатуры и сокращения, маркируя их морфологический тип (именные, наречные), а также род, число и падеж. Класс «Residual» зарезервирован для разного рода неопознанных случаев. Как видно из таблицы 1, число атрибутов варьируется от 0 (для «Residual») до 10 (для «Verb»). Таблица 2 дает представление о распределении атрибутов по частеречным классам.

Табл. 1. http://corpus.leeds.ac.uk/mocky/back.1_div.1.html)

Name (en)	Code (en)	Attributes
Noun	N	6
Verb	V	10
Adjective	A	6
Pronoun	P	7
Adverb	R	1
Adposition	S	3
Conjunction	C	4
Numeral	M	6
Particle	Q	1
Interjection	I	1
Abbreviation	Y	4
Residual	X	0

Табл 2. http://corpus.leeds.ac.uk/mocky/back.1_div.2.html)

Category (en)	Attribute (en)	Position
Abbreviation	Case	4
Abbreviation	Gender	2
Abbreviation	Number	3
Abbreviation	Syntactic_Type	1
Adjective	Case	5
Adjective	Definiteness	6
Adjective	Degree	2
Adjective	Gender	3
Adjective	Number	4
Adjective	Type	1
Adposition	Case	3
Adposition	Formation	2
Adposition	Type	1
Adverb	Degree	1
Conjunction	Coord_Type	3
Conjunction	Formation	2
Conjunction	Sub_Type	4
Conjunction	Type	1
Interjection	Formation	2
Noun	Animate	5
Noun	Case	4
Noun	Case2	6
Noun	Gender	2
Noun	Number	3
Noun	Type	1
Numeral	Animate	6
Numeral	Case	4
Numeral	Form	5
Numeral	Gender	2
Numeral	Number	3
Numeral	Type	1
Particle	Formation	1
Pronoun	Animate	7
Pronoun	Case	5
Pronoun	Gender	3
Pronoun	Number	4
Pronoun	Person	2
Pronoun	Syntactic_Type	6
Pronoun	Type	1
Verb	Aspect	9
Verb	Case	10
Verb	Definiteness	8
Verb	Gender	6
Verb	Number	5
Verb	Person	4
Verb	Tense	3
Verb	Type	1
Verb	VForm	2
Verb	Voice	7

Для корпуса ХАНКО эта спецификация была незначительно изменена в силу более детального представления морфологии в этом корпусе. Самым существенным изменением стало исключение класса «Abbreviation». Однако соответствующие атрибуты добавляются в другие части речи: акронимы для существительных (*США*) и сокращения для существительных (*г*[*од*]), глаголов (*см.*[*отри*]), прилагательных (*проч.*[*ее*]), местоимений (до *н.*[*ашей*] эры) и числительных (*тыс.*[*яча*]).. Кроме этого, были добавлены некоторые атрибуты, например, Pluralia tantum, дробные числительные и условное наклонение. Подробную информацию о ХАНКО в формате MTE можно найти по адресу **www.ling.helsinki.fi/projects/hanco/mte**.

2. Использование формата MTE для поиска ошибок аннотирования

Формат MTE обладает рядом особенностей, делающих его перспективным инструментом анализа собственно морфологических явлений и оценки аккуратности разметки. Преимуществом такого формата являются следующее.

1. **Простота и компактность**. Корпус в формате MTE представляет собой простой текстовый файл и в этом виде не требует специальных программ и оболочек. Для поиска и обработки данных можно использовать стандартные средства (например, команды семейства grep в Unix).
2. **Кроссязычность**. Стандарт MTE изначально задумывался таким образом, чтобы варианты тегсетов для разных языков были максимально близки с формальной стороны. Конечно, это невозможно соблюсти в полной мере (ср., например, количество падежей в разных языках или наличие полной формы прилагательного в русском языке). Однако в целом позиция

сходных атрибутов в тегсете и теги для совпадающих значений одинаковы для всех языков (например, **n** в пятой позиции для номинатива существительных). Эта особенность дает возможность проводить корпусные исследования по сопоставительной морфологии.

3. **Частотный анализ тегсетов**. Формат MTE позволяет анализировать не только частотность отдельных граммем, как это сделано в работах (Josselson 1953; Greenberg 1974; Копотев 2008; серия статей в (Корпусные исследования 2009) и др.), но и сочетаний граммем: целых тегсетов или их фрагментов[1].

Однако кроме этих особенностей формат MTE позволяет использовать его еще для одной процедуры, а именно для поиска ошибок в корпусе[2]. Рабочая гипотеза, которая лежит в основе этой процедуры, состоит в следующем:

> тегсеты с низкой частотой содержат большее количество ошибок, чем тегсеты с высокой частотой, поскольку хвосты частотного распределения могут быть вызваны конфликтным сочетанием тегов.

Конечно, это не значит, что все уникальные тегсеты ошибочны. Кроме того, метод не может уловить все ошибки в корпусе, поскольку теггер, с помощью которого размечался корпус, мог быть неправильно настроен и выдавать, соответственно большой массив неправильных аннотаций (известный пример такого рода — лексема *Пу́тина,* разобранная как *пути́на*). Еще одно ограничение связано с зависимым приписыванием тегов. Так, признаки падежа и числа для существительного, извлеченные из флексии, очевидно, не будут конфликты, хотя и могут быть ошибочными (например, *Людмила Пу́тина*, разобранная как форма Род. пад. Ед. ч.). Представляется, что предложенный метод особенно эффективен при оценке качества ручной обработки корпуса после автоматического аннотирования. С помощью этого метода можно частично исправить или хотя бы оценить степень аккуратности аннотирования. Для настоящего доклада была проанализирована верность/ошибочность всех тегсетов с низкой частотой для глагола, существительного и прилагательного в корпусе ХАНКО. Подсчет проводился до тех пор, пока все тегсеты, входящие в ранг, не оказывались правильными. Для существительных и глаголов это тегсеты с частотой 3, для прилагательных — 4.

2.1. Глаголы

На графике 1. показано распределение всех глагольных тегсетов.

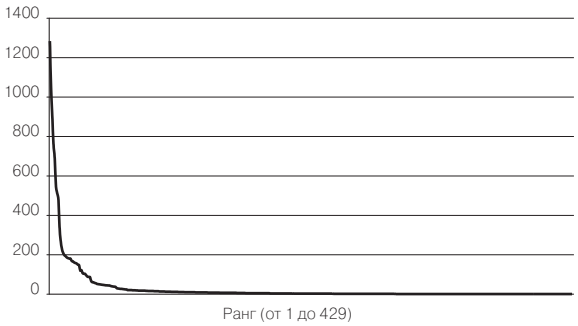


График 1. Глагол. Распределение тегсетов

Зона хвостов (с частотностью менее 10) отдельно показана на графике 2, на котором можно видеть длинный ряд тегсетов, имеющих последний ранг.

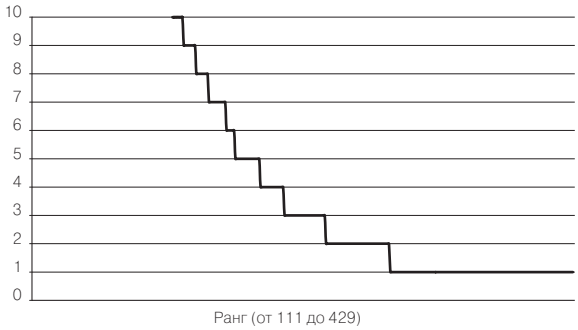


График 2. Глагол. Распределение низкочастотных тегсетов

Всего в корпусе встретилось 145 глагольных тегсетов с частотой 1. Из них 82 (56.6 %) оказались ошибочными. Приведу примеры.

1. замирившимися V-ps-**pm**afeir- замириться
   Ошибочно определен род причастия в форме множественного числа.
2. посрывало V--s3sna-e-u- посрывать
   Не определены форма и время глагола (вторая и третья позиции).
3. приводится V-ip-s-p-p-u- приводить
   Пропущен тег лица (пятая позиция).

Любопытны также и правильно определенные тегсеты. Часто достаточно одного редко встречающегося морфологического параметра, чтобы весь тегсет стал уникальным. Так, в ХАНКО последовательно маркируются двувидовые глаголы. Посколь-

[1] Близкий к этому подход разрабатывается в (Arppe 2001, Janda & Lyashevskaya [в печати]), которые используют термин «grammatical profile» для характеристики частотного распределения морфологических форм лексемы.

[2] Другие методы оценки корпуса и поиска ошибок предложены в (Brants 1995; Dickinson & Meurers 2003; Pîrvan & Tufiş 2006 и особенно Dickinson 2003).

ку они в целом достаточно редки, количество тегсетов двувидовых глаголов с частотой 1 довольно значительно.

4. использовалось V-is-snp-**b**-u- использовать
5. регламентирующих V-pp-p-af**b**g-- регламентировать

Еще одним атрибутом, влияющим на частотность всего тегсета, является сокращение (s в 13 позиции):

6. Н. V-p--p-pf-gu**s** называемый

Среди глагольных тегсетов, встретившихся в корпусе два раза, количество ошибок резко падает и составляет всего 2 %. Среди тегсетов с частотой 3 ошибочных не встретилось совсем. График 3 показывает соотношение количества ошибочных чтений для низкочастотной глаголов.

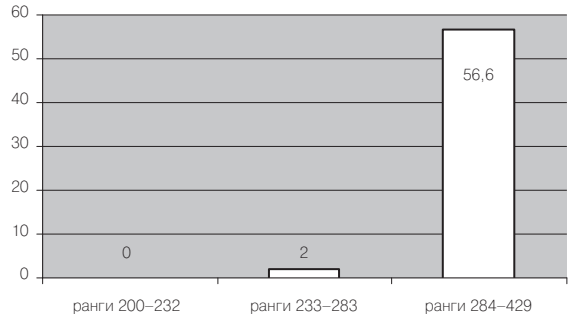


График 3. Глагол. % ошибочных тэгсетов

2.2. Существительные

На графике 4 показано распределение тегсетов существительных, встретившихся в корпусе 10 и менее раз.

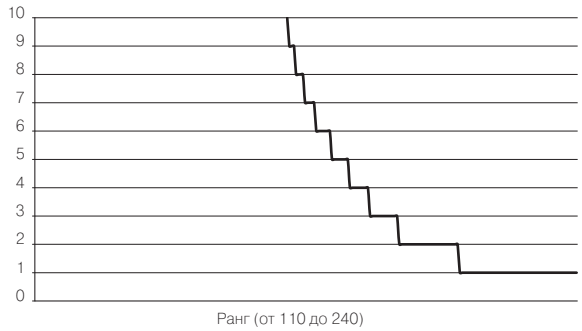


График 4. Существительное. Распределение низкочастотных тегсетов

В корпусе зафиксировано 52 тега с частотой 1, из которых 20 ошибочных (38,46 %).

Это ошибки такого рода.

7. ВЭФ N-----а ВЭФ
   Часть тегов не отмечено.
8. интернет-компаниях Nffpl--- интернет-компания
   Смешана разметка существительного (компания) и прилагательного (интернет-).
9. Лефортовской **N**-fsl--- Лефортовский
   Неверно определена часть речи.
10. Абрамович N-m-ny-- Абрамович
    Часть тегов не отмечено.

Самыми частыми атрибутами, приводящими к резкому снижению частотности тегсета, являются теги, маркирующие сокращение и акронимы:

11. ГЭС Ncfsin-**a** ГЭС
12. ул. Ncfsnn-**s** улица

а также сочетание тегов «собственное» (**p** во второй позиции) и «множественное число» (**p** в четвертой позиции):

13. Сезары N**p**m**p**an-- Сезар
14. Людвигов N**p**m**p**gn-- Людвиг

График 5 показывает соотношение ошибочных тегсетов в трех самых низкочастотных группах. Из него видно, что тегсеты с частотой 3 не содержат ошибок.

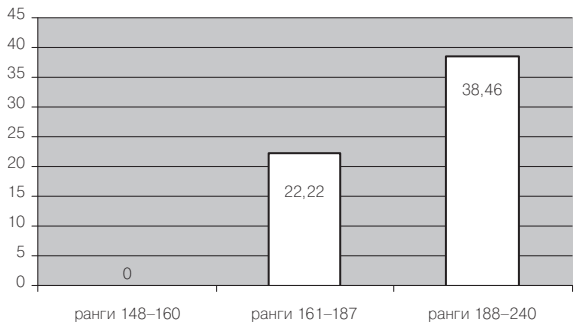


График 5. Существительное. % ошибочных тэгсетов

2.3. Прилагательные

На графике 6 показано распределение тегсетов прилагательных с частотой меньше 10.

Всего в корпусе встретилось 50 уникальных тегсетов для прилагательных, из которых ровно половина оказалась ошибочной. Часть из них — по причине неверного сочетания тегов:

15. дедушкиных As-**fp**gf- дедушкин
    Сочетание женского рода и множественного числа.

В значительной части ошибочных тегсетов не содержится тега, определяющего тип прилагательного:

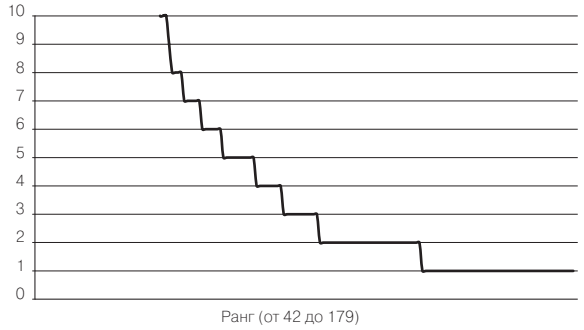


График 6. Прилагательное. Распределение низкочастотных тегсетов

16. интерросовского A--msg-- интерросовский
    Пропущен тег f во второй позиции.

Еще одна часть ошибочных чтений не содержит тега полной/краткой формы:

17. лучшему Afsnsd-- хороший
    Пропущен тег f в седьмой позиции.

Если говорить о правильных, но редко встречающихся тегсетах, то, как и для существительного и глагола, они часто вызваны сокращением слова:

18. Солнечн. Afpfsnfs солнечный
    Тег s в девятой позиции.

Ожидаемо редки притяжательные прилагательные, что приводит к снижению частотности всего тегсета.

19. Божим As-msi-- Божий
    Тег s во второй позиции.

На графике 7, как и в предыдущих случаях, показана доля ошибочно размеченных тегсетов в самых низкочастотных случаях.

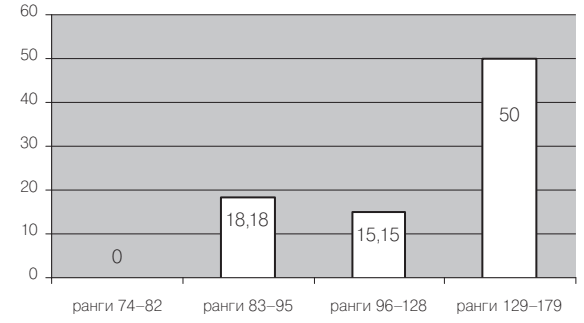


График 7. Прилагательное. % ошибочных тэгсетов

На графике видно, что для прилагательных картина распределения ошибок несколько иная: большой процент содержат не только две самых низкочастотных группы (частотностью 1 и 2), но и третья (частотность 3, ранги 83–95). Большинство из этих ошибок связано с неверным сочетанием тегов мн. числа и рода. Лишь четвертая с конца группа не содержит ошибок.

3. Выводы

График 8 дает общее представление о количестве тегсетов для трех рассмотренных частей речи (в скобках приведены абсолютные данные). Он показывает совокупную долю всех найденных ошибок к общему количеству тегсетов той или иной части речи.

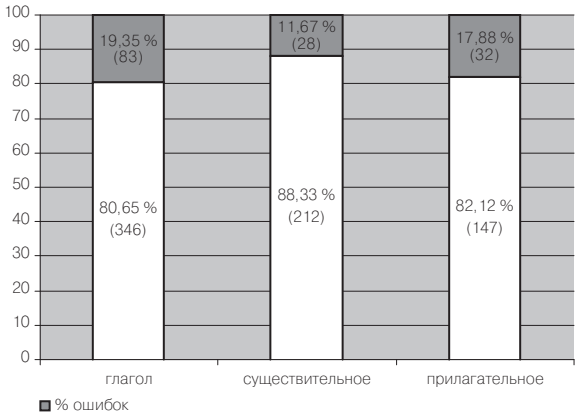


График 8. Совокупный % ошибочных тэгсетов

Представляется, что анализ низкочастотных тегсетов является полезным методом поиска ошибок, оценки и улучшения морфологической разметки русского корпуса. На примере трех частей речи удалось показать, что в зону хвостов попадает непропорционально много ошибочных тегсетов. Представляется, что на более объемном корпусе эти результаты будут еще более контрастными, поскольку частотность правильных, но редких тегсетов увеличится. При этом надо учитывать, что на частотность тегсетов влияет и схема аннотирования: чем она грубее, тем легче ее применить и тем меньше правильных низкочастотных тегсетов окажется в корпусе. Наконец, надо заметить, что соотношение ранга и частотности, установленное в свое время Джорджем Ципфом для распределения лексем, в целом верно и для морфосинтаксических тегсетов, хотя эта параллель еще требует дальнейшего уточнения и обоснования.

Литература

1. *Arppe A.* Focal points in frequency profiles — how some word forms in a paradigm are more significant than others in Finnish // Proceedings of the 6th Conference on Computational Lexicography and Corpus Research, June 28–30, 2001, University of Birmingham, Birmingham, 2001.
2. *Brants Th.* Tagset Reduction Without Information Loss // Proceedings of the 33rd Annual Meeting of the ACL. Cambridge, MA. 1995. P. 287–289.
3. *Dickinson M, Meurers, D.* Detecting Errors in Part-of-Speech Annotation // Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL-03). Budapest, Hungary, 2003.
4. *Dickinson, M.* Error detection and correction in annotated corpora. PhD thesis. The Ohio State University, 2003. Доступно по адресу: etd.ohiolink.edu/view.cgi?osu1123788552.
5. *Erjavec. T.* MULTEXT-East Version 4: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. // Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC 2010), Malta. [В печати].
6. *Greenberg J. H.* The relation of frequency to semantic feature in a case language (Russian) // K. Denning and S. Kemmer (eds), On Language, Selected Writings of Joseph H. Greenberg, Stanford, 1990 (1974). P. 207–226.
7. *Janda L., Lyashevskaya O.* Grammatical profiles and the interaction of the lexicon with aspect, tense, and mood in Russian [в печати].
8. *Josselson H. H.* Подсчет ходовых слов русского языка, Detroit (MI), 1953.
9. *Multext-East* Home Page. [Интернет-ресурс. Доступен по адресу: nl.ijs.si/ME]
10. *Pîrvan, F. Tufiş. D.* Tagsets Mapping and Statistical Training Data Cleaning-up // Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006), Genoa, May 2006. Genoa. P. 385–390.
11. *Sharoff, S, Kopotev, M., Erjavec, T, Feldman A., Divjak, D.* Designing and evaluating Russian tagset // Proceedings Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2008), Marrakech, May, 2008. Доступен по адресу: www.lrec-conf.org/proceedings/lrec2008/summaries/78.html.
12. *Корпусные* исследования по русской грамматике, Москва: Пробел-2000, 2009.

Имитация компьютерным агентом непрерывного эмоционального коммуникативного поведения

Continuous simulation of emotional communicative behavior by a computer agent

Котов А. А. (kotov@harpia.ru)

Институт лингвистики РГГУ, Москва, Россия

Мы представляем архитектуру эмоционального компьютерного агента, который реагирует на входящие компоненты семантического представления и непрерывно демонстрирует разнообразные реакции: высказывания, жесты и коммуникативные действия — почёсывания, переминания, изменения направления взгляда и т. д.

1. Введение

Традиционно лингвистическая система поддержания диалога описывается по аналогии с вычислительной машиной как система со входом и выходом, принимающая на вход высказывание пользователя и отвечающая сконструированным (или выбранным из базы) высказыванием. Если система ожидает ввода пользователя или если она занята обработкой входа, то она никак не взаимодействует с пользователем; компьютер может обозначать эти ситуации мигающим курсором (при ожидании ввода пользователя) или курсором в виде часов (если система занята обработкой входа). Таким образом, средства интерфейса для поддержания непрерывного взаимодействия с пользователем у традиционных компьютеров и лингвистических процессоров — достаточно ограничены.

Намного более жесткие требования к средствам взаимодействия с пользователем предъявляются при создании компьютерных агентов — трёхмерных анимированных персонажей (например, героев компьютерных игр) или перспективных бытовых роботов, взаимодействующих с человеком. Компьютерный агент должен постоянно выполнять разнообразные действия (почёсываться, переминаться, переводить взгляд) и при этом имитировать постоянно меняющиеся эмоциональные состояния.

Если агент взаимодействует с пользователем, то он должен комбинировать такие поведенческие реакции с речевым диалогом: одновременно говорить, жестикулировать, переводить взгляд и переминаться. В коммуникации значимыми элементами становятся достаточно простые поведенческие действия: если в обычной ситуации человек почёсывается или облизывается, чтобы устранить соответствующее раздражение (зуд, сухость губ), то в коммуникации те же действия могут передавать адресату информацию о текущем состоянии адресанта — человек может почёсываться или облизываться, неявно обозначая своё смущение. Возможность передавать информацию в коммуникации сближает такие поведенческие действия с произвольными жестами. Совокупность высказываний, а также жестов, мимики и других действий, которые могут передавать адресату информацию о внутренних переживаниях или мыслях адресанта мы будем называть *коммуникативным поведением*.

В данной работе мы демонстрируем компьютерную модель, предназначенную для имитации разнообразного коммуникативного поведения при взаимодействии с пользователем. Эта модель реагирует на входящие смыслы или события, описанные в виде простых семантических деревьев, и на выходе анимирует трёхмерную виртуальную фигуру агента.

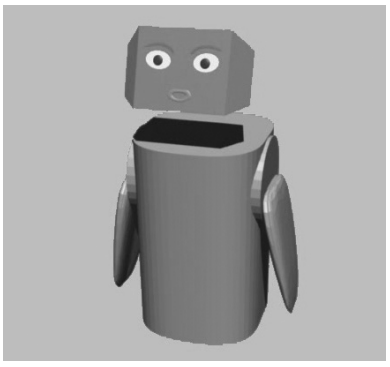


Рис. 1. Трёхмерный компьютерный агент

Для обработки входящего стимула модель использует множество параллельных альтернативных правил (сценариев), которые активизируются компонентами входа и порождают множество жестов, действий и высказываний — потенциальных кандидатов на исполнение агентом. В этой работе мы рассматриваем следующие задачи: (а) отбор жестов, действий и высказываний для выполнения, (б) выстраивание жестов и высказываний во времени (*сериализация* элементов коммуникативного поведения), и (в) исполнение выбранных жестов и действий во времени с помощью фигуры агента.

2. Трёхмерные анимированные компьютерные агенты

При анимировании трёхмерных фигур в компьютерных играх и некоторых типах интерфейсов большую важность представляет описание действий в состоянии покоя (idle motions). Если агент выполняет активные движения, то они отвлекают внимание зрителя, и мелкие действия оказываются не так важны, но в состоянии покоя полная обездвиженность или «зацикленность» движений агента выглядят неестественно или даже пугающе. Для создания правдоподобной анимации в этом случае требуется накладывать друг на друга действия трёх разных типов: (а) изменение позы, (б) повторяющиеся движения: дыхание и поддержание равновесия, (в) дополнительные действия и жесты: человек может трогать рукой волосы и лицо, класть руку в карман и т. д. [Egges, A., Visser et al., 2004]. Изменение позы — это переход между двумя относительно неподвижными позами; если расклассифицировать позы и описать для каждой пары поз разные переходы (отличающиеся выражаемой эмоций), то трёхмерная фигура получит возможность выражать в состоянии покоя заданную эмоцию, например, переминаясь с ноги на ногу более эмоционально или более сдержано [Egges, Arjan, Papagiannakis et al., 2007]. Для управления такими внешними изменениями компьютерные агенты могут снабжаться алгоритмами, имитирующими динамику эмоций, как, например, агент Макс [Jung, Kopp, 2003]. Макс «живёт» на экране в фойе Университета Билефельда и приветствует входящих людей: машет им рукой и представляется. Алгоритм управления эмоциями меняет состояние Макса в зависимости от окружающих событий. Например, отсутствие посетителей в течение длительного времени усиливает состояние «скуки», что проявляется в поведении: Макс уныло наклоняется вперёд, глубоко вздыхает, чешет голову, оглядывается, произносит *Никого!* и иногда даже уходит с экрана. Таким образом, внешнее событие меняет эмоциональное состояние, что проявляется в изменении набора жестов.

Сложные системы «эмоциональной динамики» позволяют агенту одновременно имитировать и выражать сразу несколько эмоций. Например, в рамках разработки агента Грета [Ochs, Niewiadomski et al., 2005] сделана попытка разделить испытываемую эмоцию (elicited-emotion) и выражаемую эмоцию (expressed-emotion). Грета совмещает выразительные средства этих двух эмоций при управлении мимикой, что со стороны выглядит как сложная эмоциональная экспрессия. Например, Грета может «испытывать отчаяние», при этом стараясь скрыть его с помощью внешнего проявления гнева.

Для указанных проектов сделаны расширения, где фигуры компьютерных агентов управляются традиционной системой поддержания диалога, которая принимает на вход высказывание от пользователя и передает для синтеза текст ответа, сопровождая его тэгами для управления интонацией и жестами. Если же входящего высказывания нет (система не находится в диалоге) — она выполняет движения, характерные для состояния бездействия.

3. Использование мультимодальных корпусов

Для имитации правдоподобного поведения в коммуникации мы в своей работе используем Русскоязычный эмоциональный корпус — REC [Котов, 2009] [1], где собраны видеофрагменты взаимодействия студентов с преподавателями на экзаменах, а также видеозаписи общения с клиентами в центре коммунальных платежей. Нас, прежде всего, интересуют случаи, когда человек демонстрирует разные (меняющиеся во времени) коммуникативные реакции на одно входящее событие. Такое поведение наиболее отчётливо можно наблюдать в «сильных» эмоциональных ситуациях, когда человек демонстрирует речевые реакции и сопровождает их отчётливыми эмоциональными жестами и действиями. Например, сначала ругается, потом выражает сожаление, пытается рационально решить проблему и т. д. Вместе с тем, аналогичное поведение характерно и для «слабых» эмоциональных ситуаций, когда человек использует в своём коммуникативном поведении слабовыраженные жесты или действия: чуть заметно кивает, сжимает губы, вздыхает, переводит взгляд. С помощью «слабых» жестов человек может демонстрировать аналогичную эмоциональную динамику, последовательно проявляя скрытую агрессию, сожаление и размышления о рациональном решении.

Если пронаблюдать в корпусе и смоделировать сложные «слабые» и «сильные» реакции агента на входящий стимул, то это позволит анимировать сложное эмоциональное поведение агента в тече-

[1] См. http://www.harpia.ru/rec/

ние нескольких секунд после поступления стимула (пока агент перебирает реакции на этот стимул). Если же постоянно поставлять на вход агента различные слабые стимулы и заставить агента демонстрировать сложные слабые реакции на эти стимулы (переключаясь между стимулами и между реакциями), то это позволит имитировать сложное поведение агента во времени. Агент будет почёсываться, облизываться, переминаться, рассматривать разные окружающие объекты, делать вид, что он о чём-то задумался, потом — что он что-то придумал и «принял решение» и т. д.

Одна из задач здесь состоит в том, чтобы собрать по корпусу и описать на некотором формальном языке используемые жесты и действия, а другая задача — описать и смоделировать параметры переключения между разными жестами.

4. Описание жестов и коммуникативных действий

Для описания жестов и действий в компьютерной модели мы используем язык BML (Behavior markup language) [Kopp, Krenn et al., 2006; Vilhjálmsson, Cantelmo et al., 2007]. Альтернативами BML является используемый агентом Максом язык MURML [Kranstedt, Kopp et al., 2002], языки BEAT (Behavior Expression Animation Toolkit), APML (Affective Presentation Markup Language), RRL (Rich Representation Language), VHML (Virtual Human Markup Language), а также несколько языков, предназначенных для описания объектов 3D-редакторов или элементов виртуальной реальности — такие языки менее пригодны для наших целей.

BML — это язык, основанный на XML. Он позволяет управлять отдельными элементами фигуры агента:

(111)
```
<bml>
  <head id=”h1” type=”nod” amount=”0.4”/>
  <face id=”f1” type=”eyebrows” amount=”1.0”/>
</bml>
```

или перемещать взгляд агента, то есть выполнять действие, ориентированные относительно окружающих объектов, что мы используем в нашей системе:

(112)
```
<bml>
  <gaze target="PERSON1"/>
  <speech> Hello! </speech>
</bml>
```

BML обладает системой точек синхронизации: для элементов жеста можно указывать фазы экскурсии, пика выполнения и рекурсии — и затем синхронизировать элементы жеста по этим фазам и точкам. Жест можно растягивать или сокращать во времени, смещать пик жеста, привязывая его к пикам других (сопутствующих) жестов, при этом благодаря точкам синхронизации различные элементы жеста выполняются слитно, и жест не «разваливается».

В настоящее время ведутся работы над спецификациями языков, занимающих более абстрактный уровень по отношению к BML — это язык разметки коммуникативных намерений Function Markup Language (FML) [Heylen, Kopp et al., 2008] и язык разметки эмоциональных состояний [Schröder, Devillers et al., 2007]. В нашем случае использование более высоких языков избыточно, поскольку функции имитации эмоциональных состояний и часть функций, относящихся к коммуникативным намерениям, уже успешно выполняются аппаратом сценариев.

5. Структура компонентов агента

Мы используем компьютерную модель, которая на входе принимает события в виде простых (двухуровневых) семантических деревьев и на выходе «анимирует» компьютерного персонажа. Семантическое дерево состоит из предиката и набора актантов, причём каждый из этих элементов является множеством из признаков с переменным значением [Котов, 2008]. Ранее мы использовали эту систему, чтобы приводить в движение двумерных агентов и имитировать сложные речевые реакции на входящий стимул [там же]. В данной реализации система дополнена жестовым поведением и с помощью BML анимирует уже трёхмерного персонажа, что позволяет исполнять более сложные жесты и действия, ориентированные в пространстве. В своём поведении трёхмерный агент ориентируется относительно трёх окружающих точек — это (а) лицо собеседника, (б) лицо присутствующего человека, которому агент может адресовать свои комментарии, и (в) объект действий агента, например, это может быть кнопка, которую агента просят нажать.

Система обладает следующей архитектурой (см. Рис. 2). В основе системы лежит менеджер реакций (аналогичный описанному в [Sloman, 2001]), который состоит из (а) блока входа, (б) набора сценариев для обработки входа и (в) блока выхода. В блоке выхода поддерживается трехмерная модель окружения агента (направления на ключевые точки), а также набор семантических предикатов — двухуровневых семантических деревьев, которые описывают представление агента об окружающей ситуации. Мы передаём агенту на вход семантические деревья (1), рассчитывая, что в будущем задачу извлечения семантических деревьев из текста сможет взять на себя парсер — в этом случае мы сможем передавать агенту фразы на естественном языке.

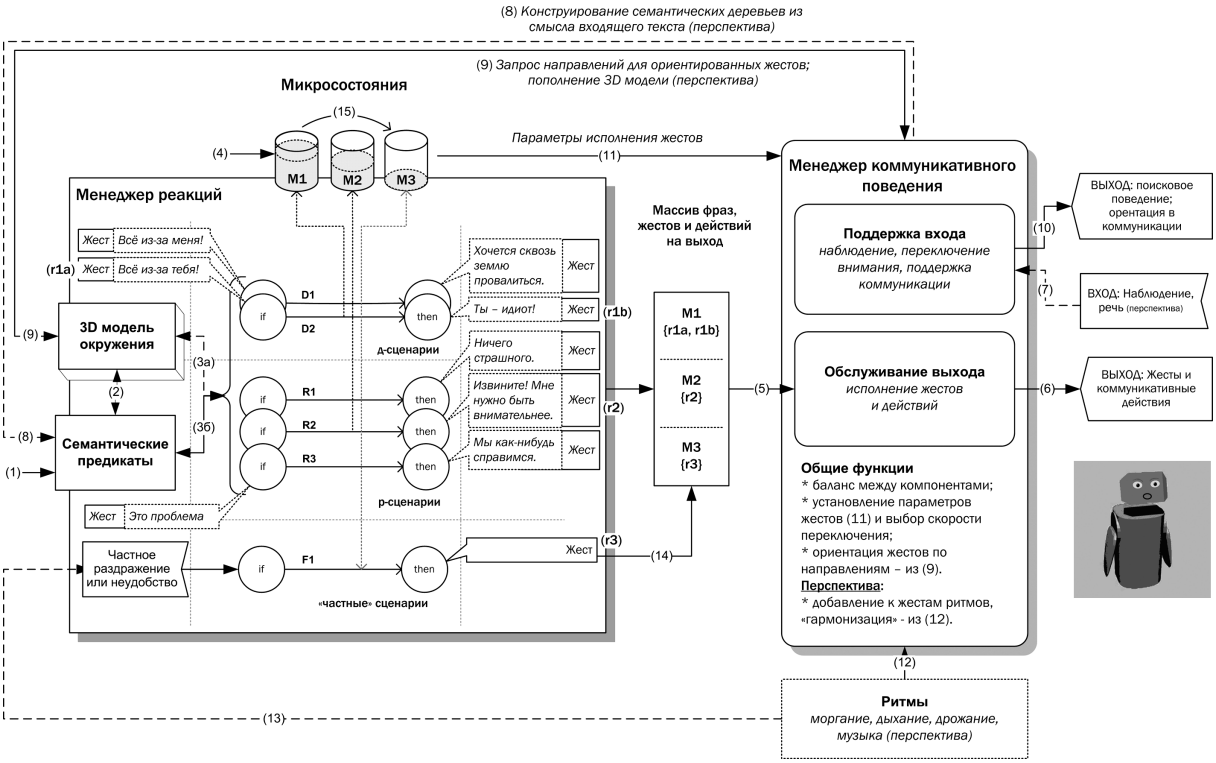


Рис. 2. Структура компонентов агента для выполнения жестового поведения (перспективные связи отмечены пунктиром)

Один и тот же объект окружающего мира может быть отмечен как референт в блоке семантических предикатов и как объект в 3D-модели окружения (2). Благодаря этому, агенту можно сообщать высказывания об определённом референте, и агент будет соотносить этот референт с точкой в пространстве, где расположен денотат. Это позволяет выполнять жесты и действия, ориентированные относительно окружающих объектов: например, если агенту сказать, что ‘он должен нажать кнопку’, то он укажет рукой на кнопку и вопросительно посмотрит на собеседника.

Входящие семантические предикаты могут активизировать сценарии (3б), каждый из которых может передать на выход высказывание и/или жест, описанный на BML. Сценарий может содержать множество альтернативных высказываний и жестов, или, к примеру, только один жест. Смежные сценарии объединены в группы и обладают общим переменным весом — этот вес содержится в *микросостояниях* (на схеме — M1–M3). Микросостояния отражают короткие эмоциональные или коммуникативные состояния, например, это ‘переживание вины’, ‘попытка извиниться’, ‘попытка показать себя с лучшей’ стороны и т. д. Меняя чувствительность микросостояний (4) можно менять «характер» или «настроение» агента, то есть те сценарии, которые будут преимущественно использоваться при обработке входа: в зависимости от настроек агент может воспринимать события преимущественно как позитивные или преимущественно как негативные и т. д. Жестовый и речевой выход, порождённый сценариями, записывается в очередь на исполнение и передаётся в менеджер коммуникативного поведения (5). При сериализации выхода в этой очереди учитываются веса микросостояний: сначала записывается и исполняется выход из наиболее активизированных микросостояний, при этом активизация микросостояний снижается, и к очереди начинают добавляться средства выражения менее активных микросостояний.

Менеджер коммуникативного поведения (МКП) исполняет жест с помощью фигуры агента (6) и управляет глазами, мимикой, руками, положением головы и тела. Для выполнения пространственно-ориентированных жестов МКП запрашивает данные о требуемых векторах у 3D-модели окружения (8). Выбранные в качестве ответов высказывания выводятся отдельно от жестов в текстовое окно на экране. На данном этапе мы пока не ставим задачу акустического синтеза речи и синхронизации между произнесением высказывания и исполнением жестов, как это сделано в системах, совмещающих в XML-разметке текст высказывания, просодию и жесты [Kopp, Wachsmuth, 2004].

При имитации поведения в МКП требуется решать множество конфликтов, главный из которых — наличие разных функций у одного и того же органа. Глаза, руки и тело могут использоваться как для обеспечения выхода (исполнения жестов), так и для

обслуживания входа (наблюдения за окружающими объектами). Например, глаза могут использоваться для наблюдения (можно переводить глаза на интересующий объект)[2] или для исполнения жеста (агент может закатывать глаза, или «бегать глазами», перемещая взгляд между окружающими объектами). Таким образом, за управление элементом «глаза» борются разные функции: задача исполнения порождённых жестов и действий (5 → 6) и задача наблюдения окружения, необходимость пополнения 3D-модели окружения (9 → 10 → 7).

По этой причине в структуре МКП выделяются два крупных блока: (а) блок обслуживания выхода — он исполняет жесты и действия, переданные из менеджера реакций и (б) блок поддержки входа — он переводит глаза на объект интереса или на собеседника (при этом агент вслед за глазами поворачивает тело и голову).

Эти блоки по очереди захватывают управление органами агента, причём баланс зависит от текущего эмоционального состояния агента (микросостояния): в некоторых ситуациях агент должен сдержанно проявлять свои жесты, но активно наблюдать за окружением, в других ситуациях — свободно и активно жестикулировать, не обращая внимания на присутствующих. Важны и параметры переключения между жестами внутри одного компонента: если в окружении агента присутствуют несколько объектов (референтов), то он может медленно переводить взгляд с одного объекта на другой или быстро бегать глазами между объектами. Действие ‘бегать глазами’ используется для выражения беспокойства агента. Это действие состоит из множества частных действий вида «переместить взгляд на новый объект», причём МКП регулирует скорость переключения между этими частными действиями в зависимости от активного микросостояния (11) — в обычной ситуации агент «осматривается» (долго фиксируется на каждом объекте), а при беспокойстве — «бегает глазами» (кратко останавливает взгляд на каждом объекте). Сейчас в качестве такого параметра передаётся скорость переключения, но в дальнейшем мы планируем передавать параметры, влияющие на скорость и амплитуду исполнения жестов, что допускается спецификацией BML: агрессивное состояние должно увеличивать скорость исполнения жеста, «зажатое» состояние — сокращать амплитуду жеста, а расслабленное состояние — уменьшать скорость и увеличивать амплитуду.

Изменение параметров выполнения жеста важно для большинства действий агента. Например, поворот головы не имеет смысла описывать для каждого сценария отдельно с разными параметрами исполнения. Поворот головы целесообразно описать в базе данных один раз и затем при исполнении накладывать на этот жест разные параметры в зависимости от текущей «эмоции» агента. То же самое относится и к частным жестам. Выразительные средства (высказывания и жесты) в базе данных связаны со сценариями отношением многие-ко-многим, то есть один жест или одно высказывание может вызываться разными сценариями (под действием разных микросостояний). Поэтому целесообразно описать каждый жест агента только один раз, а при вызове жеста из разных сценариев (микросостояний) накладывать на жест разные параметры исполнения.

В текущей версии системы МКП сглаживает переходы между смежными жестами из очереди на выполнение. Интересная перспективная задача состоит в том, чтобы выравнивать пики исполнения жестов по времени в соответствии с различными ритмами. Для этого необходимо не только растягивать или сжимать каждый отдельный жест на основании данных от микросостояний (11), но и гармонизировать смежные жесты, заставляя агента двигаться в более быстром или более расслабленном темпе или ритме. В перспективе компонент ритмов может использоваться для гармонизацией действий агента с внешней музыкой и для управления физиологическими ритмами — дыханием и морганием, которые в материале корпуса часто используются для выражения эмоциональных состояний.

6. Имитация сложных случаев коммуникативного поведения

Разрабатываемая нами система имитации коммуникативных реакций отличается от аналогичных проектов тем, что она использует простые семантические структуры для входящих событий и для представлений агента об окружающей ситуации. Это позволяет имитировать интересные формы коммуникативного поведения, связанные со сложной обработкой входа или со сложным преобразованием выхода. В области входа — это поисковое поведение, когда агенту недостаточно имеющихся данных для активизации сценария и выполнения реакции. В области выхода — это замещающие жесты, когда, к примеру, почёсывание используется для выражения состояния смущения.

6.1. Имитация различных типов поискового поведения

В ситуации агрессивной тревоги агент должен быстро оглядывать окружающее пространство. То есть поведение агента должно вызываться изменением в микросостоянии. Если мы добавим акти-

[2] В нашем случае глаза агента не используются для наблюдения, но мы считаем важным сохранить указанное разделение функций для более правдоподобной имитации поведения.

визации микросостоянию, которое объединяет сценарии типа D2 (сценарии, ответственные за агрессивное коммуникативное поведение), то это приведёт сценарии типа D2 к частичной активизации. Полная их активизация заставила бы агента кричать *Всё из-за тебя! Ты — идиот!* и махать на оппонента. Частичная активизация заставляет агента искать подтверждение «своим страхам» и пытаться обнаружить в своём окружении виновника, который мог бы занять валентность агенса в частично активизированном сценарии D2. Для этого необходимо дополнить 3D-модель окружения на основе данных, получаемых со входа. Таким образом, активизация микросостояния (4) через соответствующие сценарии приводит к запросу 3D-модели (3а), которая, в свою очередь, запрашивает блок поддержки входа в МКП (9) и вызывает поисковое поведение (10). Агент оглядывается, ища в окружающем пространстве референты, соответствующие его страхам.

6.2. Имитация замещающих жестов

Почёсывание, поправление одежды и изменение позы часто выступают в корпусе как замещающие жесты, проявляющие внутреннее волнение [Котов, 2009]. По структуре механизм замещающих жестов подобен механизму иронии, где высказывание используется для выражения скрытого коммуникативного намерения, прямое выражение которого подавлено боязнью, правилами этикета или иными причинами [Грайс, 1985; Фрейд, 1991]. Для имитации такого частного случая иронии с помощью предлагаемый компьютерной модели мы использовали механизм эксплуатации микросостояний [Kotov, 2009]. Входящая семантическая структура (1) может активизировать (3б) некоторый сценарий D2 и связанное с ним микросостояние M1 (‘агрессия, обвинения другого’), но для формирования выхода это микросостояние может обратиться (15) к другому микросостоянию M3 и связанным с ним сценариям. С помощью компьютерной системы мы имитировали случай, когда вход ‘тебя кто-то стукнул’ вызывал неудовольствие агента (активизировал негативное микросостояние M1), но при этом для ответа агент самостоятельно выбирал сценарий с наибольшей активацией из позитивного микросостояния и использовал полученное высказывание иронично: «Агент (иронично): *Хорошо, что на меня кто-то обратил внимание!*» [там же]. Механизм почёсывания в качестве замещающей реакции обладает той же структурой, но вместо замещающего высказывания агент будет использовать замещающее действие или жест.

Для обогащения поведения агента мы добавили к менеджеру сценариев множество «частных» сценариев типа F1, которые ответственны за устранение частных неудобств (почёсывание, облизывание, поправление одежды). В обычной ситуации каждый из этих сценариев активизируется своим стимулом (например, ‘у меня что-то чешется’, ‘у меня пересохли губы’), но все эти сценарии объединены общим микросостоянием — его можно условно обозначить как ‘дискомфорт’. Мы постоянно поставляем на вход агента множество частных раздражителей, обладающих сравнительно малым весом. Обычно агент начинает обрабатывать эти слабые стимулы в паузу, когда более сильные раздражители отсутствуют.

В ситуации замещающего действия активизированное микросостояние M1 начнёт эксплуатировать (15) микросостояние M3 ‘дискомфорт’, и для выражения неудовольствия агент может ‘почесаться’ или ‘облизнуться’ в зависимости от тех слабых стимулов, которые присутствуют у него на входе (M3 → F1 → 14 → 5 → 6).

Преимущество нашего подхода в том, что в ситуации раздражения агент не будет выбирать между реакциями типа ‘почесаться’ и ‘облизнуться’ случайным образом — он будет выбирать конкретное действие в зависимости от поступившего слабого стимула: ‘облизываться’, если ‘пересохли губы’, ‘почёсываться’, если ‘чешется’ и т. д. Каждое из этих действий описано в нашей системе только один раз, и при этом оно может использоваться как в своей прямой функции, так и в качестве замещающего жеста для различных микросостояний.

7. Заключение

Задача синтеза коммуникативного поведения обладает общими чертами с задачей конструирования речи. И в том и в другом случае порождённые нелинейные структуры — смысл предложения или множество разнообразных коммуникативных реакций — необходимо сериализовать и представить в виде линейной последовательности: текста или поведения. Кроме этого, различные коммуникативные реакции могут быть противопоставлены парадигматически, а при синтезе жестов и действий возникают синтагматические эффекты.

Однако между этими механизмами существует и существенное различие. При синтезе коммуникативного поведения основная роль отдаётся синтагматическому компоненту (МКП). МКП должен постоянно поддерживать правдоподобное поведение агента, используя коммуникативные реакции как материал. При синтезе поведения в «напряжённой ситуации», когда менеджер реакций порождает избыток средств выражения, мы должны пропускать лишние порождённые реакции, а в ситуации покоя, чтобы агент не простаивал, мы должны постоянно генерировать избыточное число возможных коммуникативных реакций (переминания, почёсывания, перемещения взгляда, интерес к окружа-

ющим объектам и т. д.) и заполнять ими поведение агента в реальном времени. С помощью МКП мы можем выражать эмоции и состояния агента не только с помощью отдельных жестов, но и отдавая приоритет наблюдению или экспрессии, меняя скорость переключения между отдельными объектами внимания.

Учёт компонентов семантики при выборе реакции также даёт преимущество системе синтеза коммуникативного поведения. Агент соотносит референт с объектом в пространстве и может смотреть в сторону денотата, упомянутого в речи. Кроме того, «недостаток» элементов во входящем смысле может вызывать у агента поисковое поведение, а невозможность прямо выразить свою реакцию — может заставлять демонстрировать замещающие жесты, обогащая и делая более правдоподобным искусственное коммуникативное поведение.

Литература

1. *Грайс Г. П.* Логика и речевое общение // Новое в зарубежной лингвистике. Вып. 16. Лингвистическая прагматика. М., 1985. С. 217–237.
2. *Котов А. А.* Управление динамикой речевого поведения виртуальных компьютерных агентов // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог». Вып. 7 (14). М.: РГГУ, 2008. С. 241–247.
3. *Котов А. А.* Паттерны эмоциональных коммуникативных реакций: проблемы создания корпуса и перенос на компьютерных агентов // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодн. Межд. конф. «Диалог 2009» (Бекасово, 27–31 мая 2009 г.). Вып 8 (15). М.: РГГУ, 2009. С. 211–218.
4. *Фрейд З.* Остроумие и его отношение к бессознательному // Фрейд З. «Я» и «Оно». Труды разных лет. Книга 2. Тбилиси: Мерани, 1991. С. 175–406.
5. *Egges A., Papagiannakis G., Magnenat-Thalmann N.* Presence and interaction in mixed reality environments // The Visual Computer. 2007. 23 (5). С. 317–333.
6. *Egges A., Visser R., Magnenat-Thalmann N.* Example-Based Idle Motion Synthesis in a Real-Time Application // CAPTECH Workshop. Zermatt, Switzerland, 2004. P. 13–19.
7. *Heylen D., Kopp S., Marsella S.* et al. The Next Step towards a Function Markup Language // Intelligent Virtual Agents. 2008. P. 270–280.
8. *Jung B., Kopp S.* FlurMax: An Interactive Virtual Agent for Entertaining Visitors in a Hallway // Intelligent Virtual Agents. 2003. P. 23–26.
9. *Kopp S., Krenn B., Marsella S.* et al. Towards a Common Framework for Multimodal Generation: The Behavior Markup Language // Intelligent Virtual Agents. 2006. P. 205–217.
10. *Kopp S., Wachsmuth I.* Synthesizing multimodal utterances for conversational agents // Computer animation and virtual worlds. 2004. 15. P. 39–52.
11. *Kotov A. A.* Accounting for irony and emotional oscillation in computer architectures // Proceedings of International Conference on Affective Computing and Intelligent Interaction ACII 2009. — Amsterdam: IEEE, 2009. P. 506–511.
12. *Kranstedt A., Kopp S., Wachsmuth I.* MURML: A multimodal utterance representation markup language. Technical Report 2002/05, SFB 360 Situierte Kunstliche Kommunikatoren. Universitat Bielefeld, 2002.
13. *Ochs M., Niewiadomski R., Pelachaud C.* et al. Intelligent Expressions of Emotions // J. Tao, T. Tan, and R. W. Picard (Eds.) ACII 2005, LNCS 3784. Berlin, Heidelberg: Springer-Verlag, 2005. P. 707–714.
14. *Schröder M., Devillers L., Karpouzis K.* et al. What Should a Generic Emotion Markup Language Be Able to Represent? // Affective Computing and Intelligent Interaction. 2007. С. 440–451.
15. *Sloman A.* Beyond Shallow Models of Emotion // Cognitive Processing. 2001. 2 (1). P. 177–198.
16. *Vilhjálmsson H., Cantelmo N., Cassell J.* et al. The Behavior Markup Language: Recent Developments and Challenges // Intelligent Virtual Agents. 2007. P. 99–111.

Тело в диалоге: семиотическая концептуализация тела (итоги проекта). Часть 1: тело и другие соматические объекты

Human body in a dialog: semiotic conceptualization of body (results of the project). Part 1: human body and other somatic objects

Крейдлин Г. Е. (gekr@iitp.ru)

Российский государственный гуманитарный университет, Москва

В статье подводятся итоги коллективного проекта «Части тела в русском языке и русской культуре». В ней дается формулировка и содержательная интерпретация поставленных в проекте задач, раскрывается существо применяемого в нём признакового подхода, а также обсуждается важное понятие семиотической концептуализации тела человека и связанные с ним понятия «тело» и «соматический объект». Описываются отдельные соматические объекты, их структура, функции, физические свойства, а также некоторые признаки обладателей этих объектов.

Введение. Задачи, решаемые в проекте

Человеческое общение — объект весьма сложный для изучения, и сложность эта во многом связана с тем, что, общаясь, люди пользуются не только языковыми, но и другими знаками. Изучая коммуникацию людей, важно обращать внимание не только на то, **что** одни люди говорят другим, но и **как** они это говорят. Важно понять, какими звуками и когда люди пользуются, как интонируют свою речь, как ведут себя, какие выполняют движения, как ориентируют своё тело, наконец, как они смотрят на собеседника.

Во всех перечисленных действиях ведущая роль отводится телу. Тело и его составные части, то есть телесные, или соматические объекты, а также их свойства, состояния и особенности функционирования представляют собой предмет широкого интереса со стороны не только лингвистов, но и специалистов смежных наук. Среди них выделяется невербальная семиотика — новая комплексная наука, изучающая невербальные знаки, процессы и модели невербального поведения и их использование в коммуникации. В отличие от естественных наук, лингвистика и невербальная семиотика изучают не тело человека само по себе, то есть как физический или биологический объект, а так называемую **семиотическую концептуализацию** тела[1].

Понятие семиотической концептуализации фрагмента мира, в том числе такого, как человеческое тело, представляет собой естественное расширение известного в лингвистике понятия **языковой концептуализации фрагмента мира**. Семиотическая концептуализация тела отражает представления обычного, неискушённого носителя русского языка о теле, его частях, органах и других соматических объектах в знаках сразу двух семиотических кодов — русского языка и русского языка тела (основными лексическими единицами языка тела являются жесты рук, ног, плеч и головы, мимика, знаки-взгляды, знаковые позы, телодвижения, касания и манеры поведения; в дальнейшем мы иногда будем называть просто *жестами*).

В коллективном проекте «Части тела в русском языке и русской культуре»[2] ставилась задача изучить семиотическую концептуализацию тела, то есть то, как тело представлено в знаках указанных семиотических кодов и вместе с тем сравнить их выразительные возможности. Как следствие, подготовлена база для будущего сопоставительного анализа семиотической концептуализации тела и его частей для разных кодов, этносов и культур. Такой анализ позволит, в частности, понять, как описывают форму или размер рук и ног русские, арабы или англичане, как разные вербальные и невербальные коды говорят о типовых движениях тела или какое тело считается красивым или некрасивым в разных языках и культурах.

[1] О понятии семиотической концептуализации тела и его частей см. работы Аркадьев, Крейдлин, Летучий 2008а; Крейдлин 2007; Крейдлин, Переверзева 2009а; Крейдлин, Летучий 2006.

[2] Проект был поддержан грантом РГНФ № 07-04-00203а.

Все перечисленные проблемы и их решения нашли отражение в новом, впервые предпринятом в настоящем проекте, подходе к феномену телесности — мы называем его **признаковым**. Ориентация сразу на несколько разнородных задач делает признаковый подход теоретически более адекватным и практически более удобным способом описания, чем традиционный лексикографический подход. Последний, как правило, имеет дело только с одним из рассматриваемых нами знаковых кодов и не приспособлен для их сопоставления.

Хотя теория и методология признакового подхода были изначально нацелены на решение указанных выше задач, он позволяет ответить и на другие важные исследовательские вопросы, например, как сопоставлены друг с другом отдельные семантические подсистемы, связанные с телом и телесностью. Мы можем теперь сравнивать представления в этих знаковых кодах действий, осуществляемых разными частями тела, дисфункций и аномалий, связанных с частями тела, и отвечать на вопросы следующих типов: что означают высказывания *Ноги не ходят*, *Голова раскалывается* или *Язык немеет*? Что означает, что некая часть тела сейчас или вообще плохо функционирует? Как мы об этом говорим или как это показываем? Открывается также возможность изучать механизмы и способы образования переносных значений у слов, обозначающих тело, части тела, органы и другие соматические объекты, описывать закономерности образования так называемых жестовых фразеологизмов (ср. языковые единицы *махнуть рукой на всё, потирать руки, положа руку на сердце* и соответствующие жесты **махнуть рукой**, **потирать руки** и **положить руку на грудь**) и сопоставлять все эти механизмы и способы с инструментарием невербального знакового кода.

Изучение семиотической концептуализации тела потребовало от нас решения целого ряда более конкретных, частных задач. Среди них (1) описание физических и психических свойств отдельных соматических объектов; (2) анализ имён, или номинаций, соматических объектов, прежде всего, их семантики, прагматики и синтаксического поведения, а также анализ языковых особенностей свободных и идиоматических выражений с этими номинациями. Кроме того, в ходе работы над проектом (3) были выявлены и получили объяснение некоторые особенности телесного поведения участников диалога разных тематики, жанра и стиля, прежде всего, академической лекции (см. об этом ниже); (4) охарактеризованы общие механизмы взаимодействия участников диалога, сформулированы правила контроля речевого и неречевого знакового поведения; (5) выявлены отдельные закономерности совместного функционирования русского языка и русского языка тела в коммуникативных актах разной природы и назначения.

Признаковый подход: основные характеристики

Остановимся на основных характеристиках признакового подхода, принципах и исследовательских установках, положенных в его основу. Существо признакового подхода составляет описание целого ряда множеств и их элементов. Это прежде всего (1) множества признаков и их значений, характеризующих соматические объекты; (2) множества признаков и их значений, характеризующих номинации соматических объектов; (3) множество признаков жестов, осуществляемых с активным или пассивным участием соматического объекта, и множество значений этих признаков; (4) множество признаков номинаций таких жестов и множество их значений. Кроме того, был проведен анализ (5) множеств языковых и неязыковых знаков, служащих типовыми выражениями значений каждого из признаков.[3]

Признаки вместе со своими значениями в совокупности характеризуют четыре класса единиц, которые изучались в ходе работы над проектом. Это (а) объекты мира — в нашем случае, соматические объекты, (б) объекты естественного языка — русские слова и выражения, (в) объекты языка тела — русские жесты и жестовые выражения и (г) объекты сложного, совмещённого кода — комплексные жестово-речевые формы (к таковым относятся, например, манеры поведения).

Систему признаков отличает сложная структура внутренних и внешних связей. Так, было показано, что некоторые признаки, например, цвет и ориентация, мало связаны между собой, а некоторые связаны тесно (среди них форма и размер, цвет и температура, дисфункция и способность совершать те или иные движения). В паре связанных признаков значение или изменение значения одного из них обычно имплицирует конкретное значение другого — так, высокая температура тела часто приводит к покраснению щёк.

Помимо характеристик всех классов единиц и описывающих их признаков, в семиотической концептуализации тела представлены наиболее важные свойства людей — обладателей соматических объектов, исполнителей жестов и т. д. Эти свойства могут быть разного рода. Одни относятся к разряду физических, другие — к разряду социальных, третьи принадлежат группе ментальных, четвёртые — к группе психологических, в частности эмоциональных, характеристик. Описание всех таких свойств тоже необходимо — хотя бы по той причине, что мужское тело представлено в обыденном сознании иначе, чем женское, а взрослое — иначе, чем детское. Если человек болен, то могут меняться форма, размер или

[3] О множествах, описываемых в проекте, см. подробно в работе Крейдлин, Переверзева 2010а.

цвет его тела, то есть соответствующие признаки принимают иные значения, чем у здорового человека. Известно, что цвет кожи человека может быть обусловлен расовой принадлежностью, поведение непосредственно связано с национальной и культурной принадлежностью, а форма рук и пальцев, как и свойства их поверхности, могут говорить о профессии человека. Так, мозолистые и грубые руки в представлении русских людей ассоциируются скорее с рабочими профессиями, требующими затрат физического труда, а утончённые руки, тонкие, длинные пальцы — с профессиями музыканта или художника, то есть с профессиями творческими.

Из сказанного следует, что при построении семиотической концептуализации человеческого тела и его частей и при формулировке закономерностей и правил коммуникативного поведения нам пришлось учитывать признаки обладателей соматических объектов. Это (i) пол и гендер, (ii) этнос, национальность и раса, (iii) возраст, (iv) физическое и психическое состояние, (v) социальное положение, (vi) профессия и ряд других признаков.

Классы соматических объектов

Работа над проектом велась параллельно в нескольких направлениях.

Первое из них — это построение классификации соматических объектов на основании общности их структур, функций и некоторых других свойств, определение и анализ состава каждого из классов.

Отдельный класс образует тело человека. Слово *тело* и его эквиваленты в некоторых других языках имеют несколько значений, распределённых по разным лексемам, причём число и состав этих лексем иные, чем те, что представлены в известных нам толковых словарях.

Анализ различных словарей и научных публикаций, посвящённых понятиям «тело» и «телесность»,[4] показывает, что истолковать русское слово *тело* и его аналоги в других языках через простейшие смысловые единицы (или, в терминологии А. Вежбицкой, семантические примитивы) крайне трудно, если вообще возможно. Все предпринимавшиеся ранее попытки это сделать, как убедительно показала А. Вежбицкая (Wierzbicka 2002), либо приводили к логическому кругу, либо значение слова *тело* объяснялось через такие единицы, содержание и структура которых оказывались гораздо сложнее, чем у слова *тело*. Ср., например, толкования, взятые из словаря Ожегов, Шведова 1999: *тело* — ‘организм человека или животного в его внешних, физических формах’; *организм* — ‘совокупность физических и духовных свойств человека’. В толковании слова *организм* участвуют заведомо более сложные смысловые единицы, чем ‘тело’ и даже ‘организм’, то есть более простые единицы толкуются через более сложные.

Работа в этом направлении привела учёных, в частности, представляющих наиболее крупные в Европе Московскую и Польскую семантические школы, к выводу о том, что ‘тело’ является смысловым атомом, или семантически элементарной, неразложимой единицей.

Однако слово *тело* неоднозначно, и его неоднозначность явно ощущается всеми русскоговорящими людьми.[5] Вот три предложения, свидетельствующие об этом:

(113) *Стакимтеломконкурскрасотыневыграешь*;

(114) *Егонеуклюжеетелоподдерживаликороткие толстые ноги*;

(115) *Уберите это тело с дороги!*

В примере (1) слово *тело* соответствует примитиву ‘тело’. Условно это значение (лексема ТЕЛО 1) можно описать как «остов» человека вместе с руками, ногами, головой, шеей и т. д. Ещё одно употребление этой лексемы иллюстрирует предложение *Он с наслаждением погрузил своё тело в ванну*: ведь, скорее всего, человек погружал в ванну не только «остов», но и другие части тела. В этом примере тело понимается как состоящее из ряда структурных частей, и человек представлен в анатомическом и/или физиологическом аспекте. Возможность выделить этот аспект телесности человека позволяет объяснить синонимию фраз *На операционном столе перед врачами уже лежал Ильин, покрытый простынёй* и *На операционном столе перед врачами уже лежало покрытое простынёй тело Ильина*. Дело в том, что именно в контексте врачебной деятельности, когда осуществляются манипуляции с человеческим телом, смыслы ‘Ильин’ и ‘тело Ильина’ эквиваленты, то есть слово *Ильин* интерпретируется как ‘тело Ильина’.

Человек может осмысляться не только как существо телесное, но и как как совмещающий в себе два начала — телесное и духовное. Вследствие этого в ряде контекстов слово *тело* (ТЕЛО 1) и его синонимы *телесное / физическое начало*, *телесная конструкция*, *телесная организация*[6] в современном русском языке употребляются как обозначение **физического** начала в человеке в противопоставлении

[4] Об этих понятиях см., в частности, работы Wierzbicka 2002; Рахилина 2000; Evola 2006.

[5] О многозначности слова тела см. также работы Крейдлин, Переверзева 2009б; Крейдлин, Переверзева 2010б.

[6] Эти синонимы используются исключительно в научных и художественных текстах; в быту и повседневной речи люди их не применяют.

началу **психическому**, обычно обозначаемому словами *дух*, *душа*, *душевная конструкция*, *духовность* (в одном из значений) и некоторыми другими, ср. фразы *Чтобы тело и душа были молоды*; *В здоровом теле здоровый дух*.

В примере (2) слово *тело* обозначает тот же «остов», но уже без прикреплённых к нему частей. Это значение представлено лексемой ТЕЛО 2; его можно описать как ‘тело1 без рук, ног, головы и шеи’. Лексема ТЕЛО 2 входит также в сочетание *растирание тела и конечностей* и в предложение *Сначала массажист работал над её телом, потом принялся массировать её руки и ноги*. О наличии у слова *тело* такого значения свидетельствует также сочетание *полный человек*, в котором слово *тело* вообще отсутствует. Между тем ясно, что речь здесь идёт именно о полноте туловища. Действительно, сочетание *полный человек* говорит о большом объёме тела, то есть остова, но не говорит, например, ничего о полноте рук или ног: руки и ноги у полного человека вполне могут быть худыми. Аналогично, сочетание *крепко сбитый человек*, используемое только для характеристики только мужского тела, указывает на мускулистый и твердый корпус, но не на руки, ноги, голову, шею и пр.

Лексема ТЕЛО 2 синонимична словам *туловище*, *корпус*, а также слову *торс*. Однако эти слова отличаются друг от друга смысловыми акцентами.

В слове *корпус* на передний план выходит идея тела как единства разных частей. Это проявляется, в частности, в возникновении определённых переносных значений у слова *корпус*, ср. *армейский корпус*, *корпус судна*, *корпус здания* (дома или здания, где находятся общественные учреждения, например *медицинский корпус*, *университетский корпус*).

В слове *туловище* акцентируется материальность тела. Поэтому слова *тело* и *туловище* взаимозаменяемы в сочетаниях, так сказать, «физического» характера (ср. *объём тела* и *объём туловища* — оба сочетания и допустимы и эквивалентны по смыслу), но не взаимозаменяемы в сочетаниях, говорящих о красоте (ср. *красивое тело* и **красивое туловище*).

Слово *торс* не случайно присутствует в дискурсе художников и скульпторов, то есть людей, занимающихся искусством изображения тела, а также в речи искусствоведов. Иными словами, в семантическом фокусе слова *торс* находятся эстетические характеристики тела человека.[7]

В примере (3) мы встречаемся уже с третьим значением слова *тело*, соответствующим лексеме ТЕЛО 3 ‘мертвец, труп’. Эта же лексема входит в предложения *Ворвавшись в дом, солдаты обнаружили там три тела* и *Знаешь, с охоты его принесли, / Тело у старого дуба нашли* (А. Ахматова. Сероглазый король).

Помимо указанных трёх значений, у слова *тело* есть и другие значения. Материально заполненная оболочка человеческого тела, обозначаемая лексемой ТЕЛО 4, противопоставлена незаполненной оболочке — точнее, зрительно воспринимаемому, но физически не осязаемому телесному контуру, не имеющему объёма. Последнее значение представлено в языке лексемой ТЕЛО 5. Лексемы ТЕЛО 4 и ТЕЛО 5 выступают, соответственно, в сочетаниях *упитанное тело*, *быть в теле*, *спасть с тела* (ТЕЛО 4) и *контуры тела*, *очертания тела* (ТЕЛО 5). Иными словами, ТЕЛО 4, в отличие от ТЕЛО 1, маркирует наличие у тела человека внутренних частей.

Лексема ТЕЛО 5 входит, вообще говоря, в иной синонимический ряд, чем все другие перечисленные выше лексемы, а именно в один ряд со словами *фигура, контур, силуэт*. Как обычно бывает с синонимическими рядами, лексемы этого ряда лишь квазисинонимичны, поскольку каждая обладает своей семантической (а также синтаксической) спецификой. В частности, слово *фигура* может обозначать человека, в отличие, например, от слова *контур*. Слово *силуэт* применяется к человеку, когда его тело плохо видно, например, из-за темноты. Это слово также обозначает человеческое тело в связи с необходимостью одеть тело, то есть с одеждой и, в частности, с её моделированием и шитьём (ср. название журнала «*Силуэт*», фразу *Модели тёмных оттенков с небольшим разрезом придают гармоничность силуэту*), или в связи с особенностями подачи тела в разных видах искусствах: кино, театр (театр теней), живопись.

Приведём синоптический список лексем, входящих в многозначное слово *тело*: ТЕЛО 1 — семантический примитив (в целях пояснения: ‘остов с руками, ногами, головой и шеей’); ТЕЛО 2 — ‘тело1 без рук, ног, головы и шеи’; ТЕЛО 3 — ‘мёртвый человек, мертвец, труп’; ТЕЛО 4 — ‘материально заполненная оболочка тела1’; ТЕЛО 5 — ‘материально незаполненная оболочка тела 1’.

Охарактеризовав класс соматических объектов, представленный словом *тело* в указанных значениям, кратко остановимся на других классах.

Крупные классы образуют собственно части тела (рука, нога, голова, живот и некоторые другие), части частей тела (пальцы, ладонь, подошва, голень и др.), органы и системы органов (органы пищеварения, дыхания, мочеполовые органы и др.), покровы, жидкости (кровь, желчь, слёзы), отверстия (ноздри, уши, рот, глотка) и др. Совершенно особые и интересные для исследователя классы составляют места, линии и инородные по отношению к телу соматические объекты. К местам относятся подмышка, пах, запястье, переносица, лысина, тонзура и некоторые другие соматические объекты, линии — это талия (ср. *линия талии*), линии руки, пробор, морщины и др., а инородные объекты — это наросты, прыщи, нарывы, горб, веснушки и др.

---

[7] См. обсуждение слова *торс* на сайте В. И. Беликова (http://forum.lingvo.ru/actualthread.aspx?tid=88167).

Семиотическая концептуализация различных соматических объектов

Вторая группа исследований, выполненных в проекте, была посвящена семиотической концептуализации отдельных соматических объектов. Составлены полные или близкие к полным описания таких объектов, как голова, грудь, кулак, ноги, ноздри, пальцы, плечи, пупок, руки, сердце и некоторые другие. Результаты этих описаний отражены в серии опубликованных или принятых к печати статей участников проекта[8] и представлены в серии докладов на международных, всероссийских и региональных конгрессах, конференциях и семинарах[9]. Темы, связанные с проектом, были предложены студентам РГГУ (г. Москва) и Международного университета природы, общества и человека (г. Дубна) в качестве творческих работ по курсам «Лексикография», «Теория невербальных актов», «Тело и телесность в русском языке и культуре», выполненных студентами.

При описании соматических объектов в центре внимания оказываются структурные, физические и функциональные признаки, с одной стороны, и невербальные знаковые движения, исполняемые с участием этих объектов, с другой стороны. Приведём примеры некоторых наиболее интересных свойств соматических объектов, представленных в разных языковых единицах и жестах.

Голова обладает особой мыслительной функцией, не свойственной другим соматическим объектам, а именно думает. Об этой функции головы свидетельствуют, во-первых, языковые единицы, ср. *придти в голову*, *безголовый*, *думай головой*, *У тебя своей головы нет, что ли?*, *головастый* (в этом слове представлен не только большой размер головы, но и смысл ‘умный’, то есть способный хорошо мыслить), и, во-вторых, жесты и позы, ср. мануальный жест и позу задумчивости **обхватить голову руками** и **поза мыслителя**.

Плечи выполняют разные функции, среди них по крайней мере одна является уникальной, а две другие — не уникальные, хотя свойственны они лишь малому числу соматических объектов. Первая функция — это служить подставкой для удерживания и переноски тяжёлых грузов, ср. *взвалить на плечи*; *Он посадил сына на плечи*. Другая функция плеч — связующая, а именно они связывают голову и шею с туловищем (такой же связующей функцией обладают шея и группа соматических объектов, называемых *связки*). Наконец, плечи служат границей между пространством и временем впереди и позади человека. Пространство позади человека и прошлое выражаются, например, сочетанием *за плечами* (ср. *рюкзак за плечами* и *У него за плечами тридцать лет научно-педагогического стажа*), а пространство перед человеком и будущее выражениями со словом *плечи* не обозначается (нет сочетания **перед плечами*). Ту же функцию, помимо плеч, выполняет спина, ср. *рюкзак за спиной* и *У него за спиной одни воспоминания*.

При описании внешности человека обычно обращают особое внимание на его лицо. Различают *правильные* и *неправильные черты лица*, то есть контуры и линии разных его частей. По лицу людей узнают, на лице отражаются чувства и переживания человека, а также его отношение к другому. У слова *лицо* в значении соматического объекта имеется огромное количество дериватов и переносных значений, и среди них значение ‘человек’. Подразумевается, тут, правда, человек не во всех своих ипостасях, а человек как социальная единица, то есть наделённый общественными, прежде всего гражданскими и юридическими, правами и обязанностями. *Лицо* как обозначение соматического объекта входит в целый ряд устойчивых сочетаний, часть из которых являются подлинными фразеологическими единицами, ср. *измениться в лице*, *спасть с лица*, *лицо вытянулось*, *прочесть <что-то> на лице*, ср. также номинации мимических жестов *скривить лицо* (*скривиться*), *играть лицом*, *подёрнуть лицом*. Лицо может мыслиться и как пространство (поверхность), на котором расположены многие соматические объекты самой разной природы, свойств и функций. Это органы и вместе с тем части лица (глаза, нос), это собственно части лица (например, щёки, губы), части частей лица (зрачки глаз, веки, крылья носа), отверстия (ротовое отверстие, ноздри), кости (лобная кость, челюсти). Укажем здесь лишь наиболее содержательные их характеристики.

Начнём со щёк, которые никак нельзя считать «популярным» объектом исследования среди лингвистов и специалистов по невербальной семиотике. Если глаза и уши, например, не раз служили объектом анализа, то щёки были описаны крайне плохо.[10]

В норме щёки располагаются на лице симметрично по обе стороны носа (это важная структурная характеристика щёк). Из основных физических свойств щёк мы обращаем особое внимание на три, которые не просто характеризуют щёки, но говорят многое о физическом здоровье или нездоровье человека, а также передают информацию о некоторых свойствах его тела. Речь идёт о текстуре (свойствах поверхности), форме и размере щёк.

Эти характеристики влияют на общую и эстетическую оценку внешности или физического состояния человека. Казалось бы, сочетание *жирное лицо* или *пухлое лицо* описывают лицо человека,

[8] См. работы Аркадьев, Крейдлин 2010; Аркадьев, Крейдлин, Летучий 2008б; Крейдлин, Летучий 2006; Летучий 2008.

[9] См. Крейдлин 2008; Переверзева, Крейдлин 2008; Переверзева 2009.

[10] Подробное описание щёк см. в работе Крейдлин, Летучий 2010.

однако они скорее отражают свойства щёк. Предложение ??*У него было худое лицо с пухлыми щеками*, по-видимому, не просто странное; оно ощущается как неправильное, потому что *пухлые щеки* автоматически предполагают, что лицо человека тоже является пухлым, собственно говоря, *пухлое лицо* и означает, что у человека пухлые щеки. Такая синкретичная характеристика формы и размера, как толщина, в русском языке применяется — из частей лица — не только к щекам, но также к губам и носу (*толстый нос, толстые губы*, *пухлые губы*). Однако чаще всего так характеризуются именно щёки, и в этом отношении они — наиболее «иллюстративная» часть лица. Щёки не описываются как *широкие* или *узкие*, к ним редко применяются характеристики *большие* или *маленькие*; чаще про них говорят: *пухлые*, *впалые*, *толстые, раздутые.* Большой размер щёк (их полнота) получает в русском языке дополнительные коннотации сытости и богатства человека, ср. *Во, какие щёки наел!*. Еще более показательными являются выражения малого размера щёк. Сочетания *щёки ввалились* и *впалые щёки* никогда не употребляются только для того, чтобы сказать, что у человека худое лицо: подразумевается, что либо человек бедный и потому голодный, либо его худоба — болезненная. *Одутловатые*, *вздутые, надутые, дряблые щёки* несут отрицательную оценку, а *щёчки, наливные щёки* (здесь очевидно сравнение щёк с крупными, красивыми и большими яблоками) и *круглые щёки* оцениваются скорее положительно.

Что касается «текстуры», то о щеках говорят, что они *гладкие*, *выбритые*, *свежевыбритые*, *небритые* и под., то есть покрыты или не покрыты волосами. Когда о мужчине говорят, что он *небритый*, то имеют в виду, прежде всего, актуальное состояние щёк — в большей степени, чем подбородка или шеи. Проверяют, хорошо ли человек побрился, проводя рукой по щеке, — и это делают не так, как в случае с температурой: для её проверки достаточно приложить руку ко лбу.

Щёки обладают свойством подвижности, в частности, они способны менять форму и размер. Такая способность обусловлена внутренним строением щёк, тем, что щёки — это мягкая часть лица, состоящая из мышц. Мягкость позволяет щекам *надуваться* (ср. жест **надувать щёки** и жестовый фразеологизм *надувать щёки*), а также *вздуваться, опадать, опускаться, втягиваться, расползаться* и т. п.

С физическими свойствами щёк, прежде всего с их относительно большим размером и широкой поверхностью, связано то, что щёки часто служат местом появления и нахождения на них телесных жидкостей, таких как слёзы и пот, что свидетельствует об определённом физическом или эмоциональном состоянии человека.

При исследовании различных частей тела удаётся установить один общий факт, важный для изучения синонимии языковых единиц и выявления условий перефразирования высказываний. Так, высказывания типа *Пальцы музыканта забегали по струнам* и *Руки музыканта забегали по струнам*, *У неё трепетали крылья ноздрей* и *У неё трепетали крылья носа*, то есть высказывания, отличающиеся лишь выражениями целого и его части, обычно синонимичны, поскольку части тела и части частей тела в большинстве случаев обладают большим числом общих функций и участвуют в одних и тех же действиях. Например, у ноздрей и у носа выделяются следующие общие функции, связанные с дыханием, обонянием и проникновением в организм человека объектов из внешнего мира. В примерах *Беличенко глубоко вдохнул ноздрями морозный воздух*. (Г. Бакланов. Южнее главного удара), *Ноздри стали различать уже тонкий запах гари* (П. Алешковский. Жизнеописание Хорька), *Пыль мешает ему дышать, забивается в горло и ноздри*. (Д. Рубина. Несколько торопливых слов любви) единицы *вдохнул ноздрями*, *ноздри стали различать запах* и *забивается в ноздри* можно заменить, соответственно, единицами *вдохнул носом*, *нос стал различать запах* и *забивается в нос* без изменения первоначального смысла.

Результаты анализа соматического объекта «язык» были представлены в нашей компьютерной базе данных как пример её содержательного заполнения. В этой базе отражены две основные функции языка: говорить и поглощать пищу, — а также такие важные свойства языка, как служить местом проявления болезней внутренних органов человека (ср. *налёты на языке*, *ярко-красный язык*) и показателем временной или постоянной утраты способности говорить (ср. *язык прилип к гортани*; *язык онемел*; *Прикуси язык!*). Язык может исполнять особые культурные и семиотические роли. Ср. жестокие наказания, связанные с вырыванием языка или участие языка в актах проклятья или заклинания (*Лишь только порой с языка срывались глухие проклятья*; *Типун тебе на язык!*; *Чтоб у тебя язык отсох!*), в актах передачи важной информации, ср. *пойти за языком*, где *язык* обозначает врага, который может сообщить некоторые ценные сведения. Язык выступает как показатель некоторых свойств или актуальных состояний людей. Среди них отрицательно оцениваемые свойства, такие как болтливость (ср. *язык без костей*, *болтать / чесать / молоть языком*), как неспособность хранить тайну (*длинный язык*, *развязался язык*) или не говорить того, что в данной ситуации говорить не следует (*срываться с языка*, *слетать с кончика языка*). Но это могут быть также и положительно оцениваемые свойства человека, например, остроумие и ироничность (ср. *острый язык*) либо способность хорошо говорить (*хорошо подвешенный язык*).

Шея кажется внешне однородным объектом, однако русский язык говорит нам о том, что в ней есть разные внутренние части — соматические объекты разных типов. Это косточки (кадык — только

у мужчин, шейные позвонки), жидкости (кровь, ср. *Шея налилась кровью*), жилы (*жилистая шея*), мышцы (*мускулистая шея*), жир (*заплывшая шея*; на шее могут быть жировые складки). Шея может иметь особые формы (ср. *лебединая шея*, *изгиб шеи*) и размеры (*длинная*, *высокая шея* или *короткая*, *низкая шея*). Длинная (высокая) шея обычно соотносится с худобой и неуклюжестью, ср. нормальное сочетание *худая и высокая шея* и странное *?толстая высокая шея*. Шея характеризуется также гибкостью, ср. *гибкая шея*, *вытянуть шею*. Любопытным является слово *шейка* — так иногда называют шею маленьких детей и женщин (в этом случае обозначение уменьшительно-ласкательное). Нам встретились редкие примеры со словом *шейка* применительно также к мужчинам, но в этих случаях оно обозначает малый размер шеи, ср. *Шейка у солдатика была как у балерины* (М. Веллер. Белый ослик). С шеей связано применение разного наказаний, ср. выражения *намылить шею*, *гнать взашей*, *дать по шее* Шея участвует во многих поведенческих актах и различных действиях человека, ср. *вешаться <кому-то> на шею*, *свернуть / сломать / свихнуть шею <на чём-либо>*, *гнуть шею*. Наконец, она может оцениваться эстетически, то есть как красивая или некрасивая, ср. выражения *красивая шея*, *лебединая шея*; красивую женскую шею подчёркивают одеждой, ср. *шейный вырез*.

Часть тела, называемая словом *грудь*, отличается от части женского тела, называемой *груди*. Каждое из этих слов имеет по несколько значений. Так, лексема ГРУДЬ 1 обозначает ‘верхнюю часть передней стороны туловища’, и её референт — это часть тела человека независимо от его пола и возраста. Лексема ГРУДЬ 2 обозначает одну из парных частей женского тела, а множественное число ГРУДИ 2 применяется к обеим таким частям. С другой стороны, лексема ГРУДЬ 1 также имеет обычное множественное число, ср. *При одной вести о нём вырвался вздох облегчения и радости из тысячи грудей* (Б. Савинков. Воспоминания террориста). Наконец, есть единица ГРУДИ 1, которая обозначает две груди, но, в отличие от ГРУДИ 2, не женского, а мужского тела (ср. *И сердце пьющего гражданина, ёкнувшее ранее в грудях, скатилось вовсе в пятки* (А. Росляков. Кто спит, того убьем // «Столица», 1997.11.11).

С каждой из этих частей тела связаны свои собственные функции. У части тела, обозначаемой в русском языке лексемой ГРУДЬ 1, выделяются следующие функции: (1) быть вместилищем жизненно важных внутренних органов — сердца и лёгких. Поскольку в ГРУДИ 1 находится орган дыхания, мы вполне можем говорить о такой физиологической функции стоящей за этой лексемой объекта, как (2) способность дышать, ср. *Всё так же грудь твоя легко и сладко дышит* (Ф. И. Тютчев), *дышать полной грудью*; ту же функцию косвенным образом выражают также сочетания *грудь вздымается*, *колышется*, *волнуется*. Кроме того, у лексемы ГРУДЬ 1, есть несколько функций небиологической природы. Одна из них — (3) служить защитным или атакующим средством в разного рода остро конфликтных или трудных ситуациях, ср. *встать грудью на защиту*, *выпятить / выставить / подставить грудь*, *грудью прокладывать себе дорогу в жизни* и т. п. Некоторые из этих единиц являются также номинациями жестов. (4) Культурно обусловленная функция этой лексемы связана с тем, что в груди находится сердце, — орган чувств и эмоций. Между вместилищем и вмещаемым имеется регулярная метонимия, ср. *В груди поселился страх* и *В сердце поселился страх*. Наконец, у лексемы ГРУДЬ 1 есть чётко выраженная (5) эстетическая функция; при этом эстетические оценки мужской и женской ГРУДИ 1, как правило, разные.

Женская грудь оценивается не только с точки зрения пропорционального строения, симметрии и соразмерного отношения с другими частями тела, но и с точки зрения размера, объёма, формы и внутренней структуры. Ср. в норме положительно оцениваемые *пухлая грудь*, *большая грудь* (в случае если у женщины грудь *маленькая / крохотная / крошечная* и близка по строению к мужской груди, то говорят *У неё нет груди*), *высокая / крепкая / пышная грудь* и отрицательно оцениваемые *впалая / плоская / тощая / дряблая /отвисшая грудь*. Эстетическая оценка, закрепленная за данной частью тела, иллюстрируется также сочетаниями типа *молодая грудь* и *старая грудь*; первое получает позитивную оценку, а второе — негативную. На взгляд женская грудь может быть *круглой*, *округлой*, *острой*; *обвисшей*, *дряблой*; на ощупь она может быть *твёрдой*, *мягкой*, *упругой*.

Мужская ГРУДЬ 1 в общем случае оценивается по совсем другим параметрам. Основная закономерность эстетических оценок мужской и женской ГРУДИ 1 такова: ГРУДЬ 1 мужчины и ГРУДЬ 1 женщины должны отвечать представлениям о мужском и женском начале, соответственно, и не быть похожими на *грудь*$_1$ человека противоположного пола. Если о мужчине говорят, что *у него женская грудь*, то с точки зрения красоты такая грудь, а с ней и красота самого человека, оценивается отрицательно. Если нам о мужчине говорят, что у него *широкая, крепкая грудь*, *на груди сильные мышцы*, то скрытая за этими выражениями положительная эстетическая оценка переносится на человека, и перед нашими глазами возникает образ человека физически сильного, здорового, красивого. А *впалая грудь*, напротив, говорит о мужчине физически несовершенном, слабом, болезненном.

Мужская грудь (ГРУДЬ 1) — сильная, крепкая часть тела. Отсюда, по-видимому, возникла её важная культурная функция — служить защитой от нападения на других людей, своеобразным «телесным щитом», ср. выражения *выставить вперёд грудь*,

встать грудью на защиту кого-либо или *чего-либо*. Кроме того, к такой груди *припадают*, *ищут* на ней *защиты*, *прибежища*, ср. выражение *уткнуться в грудь*. Напротив, желая продемонстрировать силу, победить человека в драке, борьбе или состязании, противника *берут за грудки*.

В русском языке есть фразеологизм, тесно связанный с формой груди — *грудь колесом*. Так говорят о мужчине — важничающем, хорохорящемся, напыщенном. Ср. *Я могу даже закрыть левый глаз, — хвастался Вася, выпячивая грудь колесом* (Д. Н. Мамин-Сибиряк. Клад).

Основная функции женской ГРУДИ 2 — биологически обусловленная и связанная с материнством и кормлением детей в самый ранний период их жизни (дети эти так и зовутся — *грудные дети*, или *груднички*). С данной функцией соотносятся такие выражения, как *давать грудь*, *прикладывать к груди*, *кормить грудью*, *сосать грудь*, *оторвать / отнять от груди*, а также глагол *вскармливать*, в значение которого *грудь*$_2$ входит в качестве встроенного актанта. *Налитая грудь* — это ‘грудь, полная молока, и потому большая’. Женская грудь, в отличие от мужской груди, таким образом, содержит выделенную часть — молочные железы. Кормление молоком и есть одно из свидетельств их наличия.

Не останавливаясь подробно на описании сердца, укажем, что о его форме, как и о форме любого внутреннего органа, мы мало что можем сказать. Впрочем, о контурах сердца мы кое-что знаем — ведь оно всегда изображается определённым образом (эта форма получила каноническое название *форма сердца*, или *сердечко*). Размер сердца тесно связан с его функциональными особенностями — так, выражение *большое сердце* говорит не столько о его размере, сколько о том, что такое сердце вмещает в себя большое количество положительных чувств, направленных на других людей (доброта, отзывчивость, щедрость, великодушие). Слово *бессердечный* указывает не на нарушение физиологической функции сердца, а именно перегонять кровь и тем самым поддерживать жизнедеятельность человеческого организма, а на важную психологическую функцию, приписываемую сердцу в русской наивной картине мира, — быть хранилищем чувств и эмоций. Ср.: *Повелевало сердце, жаждущее выразить признательность* (Ю. Давыдов. Синие тюльпаны); *Это сердце, неспособное к любви. Мрачная пустыня, а не сердце!* (И. Грекова. На испытаниях). Наконец, у сердца имеется огромное количество культурных функций (все они подробно рассмотрены в книге Уле М. Хейстада «История сердца в мировой культуре (от античности до современности)», вышедшей в Москве в 2009 году).

Заключение

Настоящая работа представляет собой первую из двух частей, в которых подводятся итоги коллективного проекта «Части тела в русском языке и русской культуре». В ней проанализирован основной объект, изучаемый в проекте, а именно понятие семиотической концептуализации тела.

При построении семиотической концептуализации тела встаёт естественный вопрос: что означает слово *тело* и из каких частей состоит человеческое тело? Отдельное место в статье была посвящена ответу на этот вопрос. Выделены отдельные значения слова *тело* и проанализирована структура его многозначности. Выявлены те смысловые элементы, без которых семантическое представление слова *тело* было бы неверным или неполным.

Помимо соматического объекта «тело», были рассмотрены и другие соматические объекты, распределённые по отдельным классам: голова, плечи, шея, грудь и др. Особое внимание было уделено признакам, описывающим их обладателей, таким как возраст, гендер, физическое состояние человека (в частности, здоров он или болен).

Литература

1. *Evola 2006 — Evola V.* St. Paul's Error: The Semantic Changes of BODY and SOUL in the Western World // Proceedings Language, Culture & Mind: Integrated Perspectives & Methodologies in the Study of Language, École Nationale Supérieure des Télécommunications (ENST): Paris, July 17–20, 2006.
2. *Wierzbicka 2002 — Wierzbicka A.* The semantics of metaphor and parable: Looking for meaning in the Gospels // Theoria at Historia Scirntarium, IV (1), 2002. P. 85 — 106.
3. *Аркадьев, Крейдлин 2010 — Аркадьев П. М., Крейдлин Г. Е.* Части тела и их функции (по данным русского языка и русского языка тела) // Сборник научных работ к 80-летию Ю. Д. Апресяна. М., 2010 (в печати)
4. *Аркадьев, Крейдлин, Летучий 2008а — Аркадьев П. М., Крейдлин Г. Е., Летучий А. Б.* Сравнительный анализ вербальных и невербальных знаковых кодов (постановка задачи и способов ее решения) // А. В. Бондарко, Г. И. Кустова, Р. И. Розина (ред.). Динамические модели. Слово. Предложение. Текст. М.: «Языки славянских культур», 2008. С. 439–449.
5. *Аркадьев, Крейдлин, Летучий 2008б — Аркадьев П. М., Крейдлин Г. Е., Летучий А. Б.* Семиотическая концептуализация тела и его частей. I. Признак «форма» // Вопросы языкознания, 2008. №6. С. 78–97.
6. *Крейдлин 2007 — Крейдлин Г. Е.* Лексикография жестов и их номинаций (словари и базы данных) // Материалы VII Международной школы-семинара «Современная лексикография: глобальные проблемы и национальные решения». Иваново, 2007. С. 17–19.
7. *Крейдлин 2008 — Kreydlin G. E.* Semiotic Conceptualization of Human Body: Lexicographical or Database System Description? // Proceedings of the 13th EURALEX. Barcelona, July 15–19, 2008. С. 702–706.
8. *Крейдлин, Летучий 2006 — Крейдлин Г. Е., Летучий А. Б.* Части тела в русском языке и в невербальных семиотических кодах // Русский язык в научном освещении, 2006, №12 (2). С. 80–115.
9. *Крейдлин, Летучий 2010 — Крейдлин Г. Е., Летучий А. Б.* Части тела в русском языке и русском языке тела. II. Щёки. М.: Русский язык в научном освещении, 2010 (в печати).
10. *Крейдлин, Переверзева 2009а — Крейдлин Г. Е., Переверзева С. И.* Признак «ориентация части тела» в семиотической картине мира // Молдован А. М. (отв. ред.) «Слово — чистое веселье». Сборник статей в честь А.Б. Пеньковского. М.: Языки славянской культуры, 2009. С. 337–349.
11. *Крейдлин, Переверзева 2009б — Крейдлин Г. Е., Переверзева С. И.* Тело и его части как объекты семиотической концептуализации // Tilman Berger, Markus Giger, Sibylle Kurt, Imke Mendoza (Hg.) Von grammatischen Kategorien und sprachlichen Weltbildern. Festschrift für Daniel Weiss zum 60. Geburtstag. München — Wien: Wiener Slawistischer Almanach, 2009. С. 369–384.
12. *Крейдлин, Переверзева 2010а — Крейдлин Г. Е., Переверзева С. И.* Семиотическая концептуализация тела и его частей. I. Структурные характеристики соматических объектов. М., 2010 (в печати).
13. *Крейдлин, Переверзева 2010б — Крейдлин Г. Е., Переверзева С. И.* Семиотическая концептуализация тела и его частей: тело, части тела и телесность // Теоретические и прикладные аспекты современной филологии: материалы XV Всероссийских филологических чтений имени проф. Р. Т. Гриб (1928–1995). Красноярск: Сибирский федеральный университет, 2010 (в печати).
14. *Летучий 2008 — Летучий А. Б.* Часть тела/форма «кулак»: функции, концептуализация, место в системе частей тела // Вестник РГГУ (Московский лингвистический журнал). М.: Издательский центр РГГУ, 2008. №6 (Т. 10). С. 91–108.
15. *Переверзева, Крейдлин 2008 — Переверзева С. И., Крейдлин Г. Е.* Телесность и некоторые особенности семиотического диалогического поведения // Компьютерная лингвистика и интеллектуальные технологии (по материалам ежегодной Международной конференции «Диалог» (Бекасово, 4–8 июня 2008 г.)). М., 2008. Вып. 7 (14). С. 427–430.
16. *Переверзева 2009 — Pereverzeva S. I.* Human body in the Russian language and culture: the features of body and body parts // Proceedings of the International Conference on Gesture and Speech in Interaction. September 24–26, 2009 (электронный ресурс)
17. *Рахилина 2000 — Рахилина Е. В.* Когнитивный анализ предметных имён. М., 2000.

Тело в диалоге: семиотическая концептуализация тела (итоги проекта). Часть 2: признаки соматических объеков и их значения

Human body in a dialog: semiotic conceptualization of body (results of the project). Part 2: features of semiotic objects and their values

Крейдлин Г. Е. (gekr@iitp.ru),
Переверзева С. И. (P_Sveta@hotmail.com)

Российский государственный гуманитарный университет, Москва

Настоящая работа является естественным продолжением статьи «Тело в диалоге: семиотическая концептуализация тела (итоги проекта). Часть 1: Тело и другие соматические объекты». В ней содержательно раскрываются два направления, в которых шла работа над проектом. Это, во-первых, построение классификации признаков, характеризующих тело человека и его части, и, во-вторых, построение компьютерной базы данных, суммирующей результаты описания семиотической концептуализации тела и телесности.

Введение

В настоящей статье, которая представляет вторую часть работы «Тело в диалоге: семиотическая концептуализация тела (итоги проекта). Часть 1: Тело и другие соматические объекты», речь идёт об исследованиях, предпринятых в рамках коллективного проекта «Части тела в русском языке и русской культуре»[1]. Эти исследования включают в себя подробное описание русской семиотической концептуализации соматических объектов, то есть представления обычных, неискушённых носителей русской культуры о теле и его различных частях — органах, мышцах, покровах, жидкостях, костях и т. д. Нас интересовало, как «наивные» представления о телесности отображаются в знаках двух семиотических кодов — русского языка и русского языка тела. В целях сопоставительного анализа выразительных возможностей каждого из этих кодов мы разработали специальный признаковый подход к описанию семиотической концептуализации тела и его частей. Суть этого подхода изложена в работах Крейдлин 2010 и Крейдлин, Переверзева 2009. Здесь мы лишь вкратце напомним, что признаковый подход предполагает выделение и детальное описание серии множеств, среди которых важное место занимают (1) множество признаков, характеризующих соматические объекты, и (2) множество значений этих признаков.

Одна из задач проекта состояла в классифицировании признаков соматических объектов и классифицировании их значений. Классификации строились с таким расчётом, чтобы с их помощью можно было в будущем решать разные научно-исследовательские задачи, прежде всего, задачу межкультурного сопоставления способов семиотической концептуализации тела и его частей и соответствующих знаковых кодов. В связи с этим кажется естественным считать множество выделенных нами признаков пока что открытым, не говоря уже о множествах значений каждого из признаков и о множествах языковых или жестовых выражений, обслуживающих такие значения. На сегодняшний день в нашей системе имеется порядка 90 признаков.

При работе над проектом были обнаружены и описаны важнейшие структурные, физические и функциональные признаки, характеризующие тело и его части. Мы ставили задачу выявить, какие значения принимает каждый признак для того или иного соматического объекта, а также проанали-

[1] Проект был поддержан грантом РГНФ № 07-04-00203а.

зировать основные способы выражения всех этих значений в русском языке и русском языке тела. Настоящая работа призвана отразить теоретические результаты этого исследования (представить основные признаки трёх указанных классов, их значения и языковые и жестовые единицы, передающие эти значения) и практические результаты (кратко описать структуру базы данных и привести примеры запросов пользователя).

1. Структурные признаки соматических объектов

К структурным признакам относятся, среди прочих, членимость соматического объекта на составные части, языковая или жестовая маркированность отдельных (более мелких) соматических объектов в составе данного объекта и внутренний состав соматического объекта, то есть характеристика множества соматических объектов, содержащихся внутри данного объекта.[2]

Описывая часть тела «рука», мы отмечаем, что она включает в себя часть от плеча до локтя, которая в русском языке не имеет отдельного идиоматического названия; как и целое, она называется *рука*. Кроме того, в состав руки входит, например, кисть, у которой есть свои составные части — ладонь и пальцы. У пальцев есть костяшки (не *кости* и не *косточки*!) и подушечки (не *подушки*!). Что же касается фаланг, то, во-первых, слово *фаланга* плохо освоено обычным русским языком, а во-вторых, фаланги практически не участвуют в производстве жестов. Иными словами, фаланги не входят в центр семиотической концептуализации тела и его частей.

При описании структурных признаков мы встретились с рядом проблемных случаев. Так, вопрос, являются ли брови составной частью глаз, однозначного решения не имеет. Обычно люди не считают брови частями глаза и нередко их противопоставляют, ср. хотя бы выражение *не в бровь, а в глаз*. Они понимают под бровями ‘две небольшие полоски волос над глазами’ и считают, что брови предохраняют глаза от некоторых внешних воздействий, а также выражают определённые мысли и чувства. Между тем специалисты, изучающие свойства глаз, движения зрачков и знаки-взгляды, часто относят брови к составным частям глаз. Например, основатель кинесики — науки о жестах, жестовых процессах и жестовых системах — американский антрополог Рэй Бирдвистел, описывая знаковые движения глаз, свойственные американцам англосаксонской культуры, рассматривал жесты бровей как подкласс жестов глаз. К жестам бровей относятся, например, мимические знаки недоумения (**поднятые вверх брови**, характерные для удивлённых или изумлённых глаз) и **поднятие вверх одной брови** (**lifted movement**, или **single-brow movement**) — мимическое выражение скепсиса.[3]

Признак «языковая или жестовая маркированность отдельных соматических объектов в составе данного объекта» выделяет одни соматические объекты как более важные для жизни и деятельности человека, чем другие. Выделение того или иного соматического объекта в составе данного предполагает членимость исходного соматического объекта на выделенный объект и оставшуюся часть, которая обычно получает то же имя, что и целое, например, в составе языка выделяется часть *кончик языка*, а остальная часть называется *язык*. Если членимость подразумевает деление соматического объекта, так сказать, без остатка, то выделенность, или маркированность, делает заметным (salient) только один или несколько объектов в составе данного, которые всегда получают специфические наименования.

Названия выделенных частей соматического объекта обладают следующим интересным свойством: они вместе с названием самого соматического объекта выступают в конструкциях *X, точнее <точнее сказать, то есть, а именно* и т. п.*> Y*, а также *Не весь X, а Y*, где X — имя соматического объекта, а Y — имя его выделенной части. Ср. *Набитую киркой землю он руками, точнее сказать, ладонями <…> сдвигает к зеву пещеры* (В. Маканин. Лаз); *Лоб основательно забинтован. Не вся голова, а именно лоб* (А. Ткачёва. Приворот).

Внутри одного соматического объекта могут находиться другие объекты, причём род (тип) этих внутренних объектов может быть разным. Например, с точки зрения русского языка и русской культуры во внутреннем составе руки (*в руке*) выделяются мускулы, кровь, вены, жир и кости, внутри языка находятся кровеносные сосуды, а внутри волос (волосяного покрова) нет никаких соматических объектов.

2. Физические признаки соматических объектов

К физическим признакам, помимо уже упоминавшихся формы и размера, относятся цвет[4], температура[5], звучания, издаваемые телом и его частями или рождающиеся в них, сухость и ряд других.

---

[2] Об этих и других структурных признаках соматических объектов см. работу Крейдлин, Переверзева 2010.

[3] См. Бирдвистел 1967.

[4] Подробнее о признаке «цвет соматических объектов» см. в работе Кадыкова, Крейдлин 2010.

[5] писание признака «температура» см. в работе Крейдлин 2009.

Температура тела тесно связана с психическими свойствами человека, прежде всего с чувствами и эмоциями, им испытываемыми. Сочетаемость температурных прилагательных со словами, обозначающими чувства и отношения, весьма прихотлива: мы говорим *горячая любовь*, но не **теплая любовь*, *холодная ярость*, но не **ледяная ярость*. Допустимыми являются как *холодный ужас,* так и *ледяной ужас*, как *холодное презрение*, так и *ледяное презрение*, но недопустимо ни **холодное возмущение,* ни **ледяное возмущение. Теплое чувство* (*теплые чувства*) — это чувство любви, дружбы, симпатии, приязни. Теплые чувства создают комфорт и уют. Они обычно проявляются невербально — через нежные, теплые, ласковые взгляды, приветственные жесты и радостные улыбки при встрече. Если *теплое тело* — это про температуру тела, то *теплый человек* — это про свойства данного человека, про то, что рядом с этим человеком испытываешь чувство комфорта и уюта. Некоторые из подобных сочетаний плохо интерпретируемы: что означают, например, такие сочетания, как *теплые волосы, горячие костяшки, холодная талия*?.

Что же касается сочетаний слова *теплый* с названиями частей тела и органов, то с одними из них *теплый* ведет себя исключительно как температурное прилагательное, а с другими — не только как температурное. Одно дело *теплое сердце* и совсем другое — *теплый лоб*, одно дело *теплые руки* и другое — *теплый нос.* Матери часто проверяют у своих маленьких детей, *теплый* ли у них *нос*, и если нос *теплый*, то это означает, что ребенок не замерз. О температуре органов, которые, как сердце, человеку в норме не видны, мы судить не можем. Потому-то и возможно только одно, эмоциональное, осмысление атрибутивных сочетаний *теплое сердце* и *теплая душа*, ср. *Простая, прямая и теплая душа его искала опоры в вере народной* (Н. Лесков. Владычный суд). А вот *теплые руки* могут говорить не только о температуре рук, но и о нежности и ласковости их обладателя: <…> *кто-то обвил его шею мягкими теплыми руками и зашептал на ухо* (Е. Богданов. Вьюга).

При изучении категории цвета мы отталкивались от тела и его частей, то есть тех объектов, к которым эта категория применима особым образом. Так, цвет тела и отдельных его частей может быть постоянным и переменным: таковы, например, цвет глаз или цвет кожи. Фраза *Её красивые зелёные глаза стали красными от слёз* абсолютно нормальна, потому что никакого противопоставления признаков здесь нет: сочетание *зелёные глаза* характеризует постоянный цвет глаз, а сочетание *красные глаза* — актуальный. *Красные глаза* — это обозначение переменной характеристики цвета глаз.

Изменение цвета может означать: (1) изменение физического состояния человека. Например, когда человек нездоров, мы говорим, что он *стал серым*, *пожелтел*, *побелел*, или что он *весь зелёный*, а в состоянии физического напряжения человек часто *краснеет*; (2) изменение его психического, в частности эмоционального, состояния, ср. *Она побледнела от злости* или *покраснела от гнева*; (3) изменение условий его существования. Под действием сильного солнца *чернеет*, *краснеет*, его кожа *становится бронзовой*, *коричневой*. Когда человек переезжает из средней полосы на север, его лицо становится менее ярким (*белым*, *серым* и т. д.).

Цвета отдельных соматических объектов могут отражать принадлежность людей к определённым социальным, этническим и даже возрастным группам. Например, цвет кожи характеризует принадлежность человека к определённой расе, ср. сочетания *белый человек* (*белый*), *темнокожий человек* (*негр*), *краснокожий* и под. Выражения *белая кожа*, *белая / чёрная кость*, *голубая кровь* обозначают то, что человек входит в тот или иной слой общества (высший свет, низшие слои). Исходно цветовые прилагательные в своих переносных значениях обозначают возраст человека (*пожелтевшее лицо* — так часто говорят о пожилых людях, *розовые пяточки* — характерная черта младенцев). Существуют цветовые обозначения тела и частей тела, которые стереотипно приписываются красивым русским девушкам и юношам (ср. *русые косы*, *белая грудь*, *белая кожа*, *голубые глаза*, *алые губы*).

Признак «сухость» принимает значения, обычно представляемые в русском языке словами *сухой*, *мокрый* и *влажный*. Они могут обозначать как актуальное состояние данного соматического объекта, так и его постоянное свойство. Например, язык всегда влажный, это его постоянное свойство, и потому сочетание *влажный язык* встречается крайне редко (в Русском национальном корпусе оно вообще отсутствует). Между тем, в русском языке есть выражение *сухой язык* (в Корпусе примерно 20 вхождений), обозначающее актуальное состояния языка.

Слова *сухой*, *влажный* и *мокрый* объединяет то, что они, во-первых, характеризуют разные соматические объекты, ср. *сухие* (*влажные*, *мокрые*) *руки*, *глаза*, *лицо*, *спина*, *кожа*, и, во-вторых, что все они могут свидетельствовать о физическом недостатке, указывать на болезнь или обозначать проявление определённых эмоций. Например, если один человек говорит другому, что у него *сухие руки*, то он сказанным обычно даёт ему понять, что руки адресата требуют особого ухода, например, увлажнения при помощи специальных кремов. *Сухие глаза* обозначают неприятную болезнь глаз, *сухотка* — это болезненная худоба, а *сухотка спинного мозга* — это название тяжёлой болезни, как и *обезвоживание организма*. Преклонному возрасту свойственны и физические недостатки, и болезни. Сочетания *сухая* или *сухощавая старушка* говорят о сухости тела в целом, недостатке в нём не только влаги, но и жизненной силы. Наконец, *высохшие глаза* могут описывать человека, лишённого каких-либо эмоций или даже желания жить дальше.

В разделе проекта, посвящённом звукам и звучаниям соматических объектов, центральную часть составляет анализ основных противопоставлений. Это (1) звуки, описываемые в вербальных единицах, и звуки, совмещаемые с жестами (ср. языковое выражение *хлопать в ладоши от радости* и жест **аплодисменты**, выражение *щёлкать косточками* и жест **щёлкать пальцами** при вспоминании слова); (2) Звуки, издаваемые соматическим объектом, то есть для которых данный объект является источником (ср. стук сердца, пульсация крови), и звуки, для которых соматический объект является лишь вместилищем (звон в ушах, клокотание в горле). Это противопоставление, однако, не бинарное — есть звуки, которые входят в оба класса, так, мы говорим *голова шумит* и *в голове шумит*; *живот урчит* и *в животе урчит*; (3) Звуки, уникальные для данного объекта, и звуки, свойственные многим объектам. К уникальным звукам относятся шарканье и топанье ног, храп, крик, скрежет зубов, чмоканье губами и др. Неуникальные звуки — стук (крови в висках, сердца), хруст (звук, характерный для разных костей); (4) Звуки как знаки (коммуникативные звуки) и звуки незнаковые как характеристики некоторых физиологических проявлений или действий. Примером первых является свист как призыв и как знак одобрения / неодобрения зрелища; примером вторых является звук от почёсывания; (5) Звуки, в норме слышимые только обладателем соматического объекта, и звуки, которые могут слышать другие люди. К первым относятся, например, шум в ушах или биение сердца, а ко вторым — стон, звуки плача, звуки икоты; (6) Звуки, сопровождающие жесты, и звуки, в норме не сопровождающие жесты, ср. звук от перебирания пальцами, сопровождающий один из жестов задумчивости и шмыганье носом, которое не связано ни с каким жестом; (7) Звуки, входящие в фонетическую систему русского языка, и параязыковые звуки русского языка тела (ср. гласные звуки русского языка и звуки кашля); (8) Одиночные (простые) звуки и сочетания звуков (комплексные звуки), ср. произнесение междометия *А!* и звуки храпа, пения, рыданий; (9) Непрерывные и прерывистые звуки (ср. шум и стук); (10) Высокие и низкие звуки (визг, писк, звон — высокие звуки; бас, урчание — низкие); (11) Звуки, связанные с конкретным семиотическим актом, и звуки, не связанные ни с одним семиотическим актом. Хныканье ребёнка является явным выражением каприза, *М-м?* — это вопрос-недоумение, горловые междометные звуки являются вокальными жестами, интерпретируемыми как ответные речевые реплики, в частности, выражение согласия или неодобрения. К звукам, не связанным с семиотическими актами, относятся, например, звуки храпа или стук зубов (от холода). Есть и другие оппозиции на множестве звуков, но мы на них здесь останавливаться не будем.

Помимо противопоставлений звуков существуют важные противопоставления на множестве языковых единиц, непосредственно связанных со звуками и звучаниями соматических объектов. Первое из них — это оппозиция «обозначение реального звука — обозначение ощущения (чаще неприятного, вплоть до болевого), представляемого через звуки», ср. сочетания *сердце стучит* и *ноги гудят*, *кости трещат* и *голова трещит* (первые обозначают реальный звук, а вторые — звук воображаемый: в их семантику входит отсылка к реальному звуку, издаваемому неким эталонным объектом).

3. Основные функции соматических объектов

Построение семиотической концептуализации тела и телесности в качестве важных шагов предполагает анализ основных функций соматических объектов, то есть их назначения, обеспечивающего жизнедеятельность человека и определяющего его поведение в разных жизненных ситуациях, а также изучение типовых нарушений нормального функционирования соматических объектов.[6] Показано, в частности, что функции частей тела согласуются с их внутренним устройством и реализуются в основном в действиях, которые эти части тела могут выполнять или над которыми они производятся. Функции обусловлены, с одной стороны, анатомическим строением человеческого тела, а с другой, — теми физиологическими и социальными задачами, которые человек решает на своём жизненном пути.

Выделяя функции соматического объекта и указывая их языковые обозначения, мы обычно выбирали для этого такие языковые единицы, которые свидетельствуют о реализации данной функции. Между тем неисполнение функции тоже является одним из её возможных проявлений, нередко не менее важным, чем собственно реализация. Например, одна из основных функций части тела «язык» — это говорение, однако важным аспектом этой функции является также молчание, проявляющееся, например, во фразеологических выражениях *держать язык за зубами*, *проглотить язык, прикусить язык* и т. п. Аналогичным образом, «не воспринимать звуки» является одной из реализаций функции *ушей*, ср. *заткнуть уши*; *У него в ушах бананы*.

Часто бывает, что о функциях люди узнают не по здоровому состоянию тела или его части, а по их патологии. К патологическим состояниям частей тела мы относим аномальное функционирование, вызванное внешними или психологическими

[6] О функциях тела и его частей см. статью Аркадьев, Крейдлин 2010.

причинами, болезнь (нездоровое состояние) и отсутствие части тела.

Выражения в языке аномального функционирования и болезней являются надёжными показателями наличия у данной части тела той или иной функции. Когда язык *немеет*, он не может выполнять свою основную функцию — говорить; когда *коченеют* или *деревенеют ноги* (например, от холода), они с трудом передвигаются. От ужаса или других сильных отрицательных эмоций *застывает взгляд*, и глаза плохо выполняют свою функцию – воспринимать зрением. Это всё примеры временного аномального функционирования разных частей тела. Препятствием для нормального функционирования соматических объектов являются также болезни. Поэтому главной целью врачебной деятельности — назначения лекарств, различных лечебных процедур, оперативного вмешательства и др. — является восстановление утраченных функций. Отсутствие части тела тоже маркируется в языке, ср. выражения *Ему оторвало ногу*; *выколоть глаз*; *У него нет пальца*. Как показывает опыт человеческой жизни, многие части тела обладают способностью при необходимости выполнять несвойственные им, «чужие» функции. При отсутствии правой руки многие её функции берёт на себя левая рука, а при отсутствии обеих рук — ноги (например, удерживать небольшие предметы и выполнять с ними некоторые действия). Зрение как способность воспринимать окружающий мир может компенсироваться слухом и тактильной чувствительностью.

В проекте построена классификация функций соматических объектов по двум важным дифференциальным признакам: (а) необходимость для физиологического существования человека (по этому признаку функции частей тела делятся на физиологические и нефизиологические); (б) обусловленность функции (в соответствии с этим признаком выделяются биологически обусловленные и небиологически обусловленные функции).

Между биологически и небиологически обусловленными функциями есть принципиальное различие. Первые характеризуют тело человека как таковое, вторые же характеризуют человека и его деятельность, его существование в социуме и культуре в связи с его телом. К физиологическим мы относим те функции соматических объектов, которые направлены на поддержание жизни и здоровья человека, целостности его организма. Так, физиологической функцией сердца является перегонять кровь, желудка — переваривать пищу, лёгких — забирать кислород из воздуха и выдыхать углекислый газ.

Остальные функции частей тела — нефизиологические. Для глаз это — видеть, для ушей — слышать, для носа и ноздрей — ощущать запах (обонять), для кожи — испытывать тактильные ощущения.

Небиологически обусловленные функции соматических объектов делятся на подклассы в зависимости от того, какие сферы человеческой деятельности они обслуживают. Так, выделяются культурно обусловленные функции. Они, не имея непосредственной биологической мотивации, возникают в силу того, что соматические объекты используются в важных культурных процессах, например, в ритуалах, в искусстве, в спорте, играх и т. д. Такие функции приписываются соматическим объектам в силу определённых культурных конвенций. Примерами культурно обусловленных функций являются участие *руки* в ритуалах невербального приветствия, ср. *рукопожатие*, *протянуть руку для поцелуя*, *отдать честь*, *объятие*, участие *глаз* в таких культурных актах, как флирт (ср. *подмигнуть, строить глазки*), или игра (*гляделки*).

Социально обусловленные функции — это функции, определяющие место и назначение соматических объектов в разного рода социальных процессах. Ср. участие руки и некоторых пальцев в актах указания (дейксиса) на человека, объект или направление, использование ноги в ряде инструментальных действий — от чисто производственных, например, приведение в действие каких-либо устройств, до эстетических, таких как игра на органе или управление педалями рояля.

Важнейшим подтипом социальных функций соматических объектов, который мы по причине его высокой значимости, выделяем в отдельный класс, являются коммуникативные функции. Коммуникативной функцией говорения обладает комплекс органов, именуемых *речевым аппаратом*. Однако и объекты, не входящие в речевой аппарат, могут выполнять коммуникативные функции. Так, лицо и его части участвуют в производстве мимических знаков, корпус и ноги — в позах и знаковых телодвижениях, а руки, ноги, плечи и голова — в жестах соответствующих частей тела.

Такой важный класс функций, как эстетические функции, связан напрямую с восприятием тела, внешности человека, его красоты или уродства. Эстетические функции соматических объектов соотносятся с такими социально значимыми категориями, как личный и общественный успех, большая или меньшая жизненная активность человека. У некоторых соматических объектов эстетическая функция является одной из основных; это черты лица, это глаза, волосы, грудь, ноги и т. д.

4. Компьютерная база данных «Семиотическая концептуализация тела»

Информация об именах соматических объектов, о жестах, исполняемых с их участием, а также о свойствах этих объектов, отражаемых в русском языке и русском языке тела, хранится в компью-

терной базе данных, суммирующей результаты описания семиотической концептуализации тела и телесности. Основной модуль базы данных образует программа поиска соматических объектов, их признаков, а также вербальных или невербальных выражений значений этих признаков. Её можно рассматривать как начальный этап в создании справочного портала, предназначенного для широкого круга пользователей и не предполагающего профессиональных знаний компьютера или языков программирования.

Программа состоит из двух подпрограмм: первая предназначена для пользователя, и с её помощью производится поиск, вторая, более техническая, специально создана для участников проекта, с её помощью исследователи могут пополнять и корректировать имеющуюся информацию о теле и телесности, о строении семиотической концептуализации тела, а также отлаживать систему в целом.

Программа, предназначенная для пользователя, позволяет получить ответы на следующие три вопроса: (1) какими признаками обладает данный соматический объект (например, какими признаками обладает часть тела «спина»)? (2) какие соматические объекты обладают тем или иным признаком (например, для каких объектов определён признак «гибкость»)? (3) какие значения принимает данный признак для данного соматического объекта (например, какие значения принимает признак «форма головы», или, менее формально, какие формы бывают у головы)? При ответе на последний вопрос пользователю выдаются также отдельные языковые единицы (слова, свободные или устойчивые сочетания, в частности номинации жестов, или фразы), выражающие значения разных признаков разных соматических объектов.

Для построения базы данных мы привлекли к работе Е. Клыгину и Э. Заришеву, студенток отделения «Интеллектуальные системы в гуманитарной сфере» Института лингвистики РГГУ. Они не только помогли нам в создании базы данных, но и получили возможность применить свои знания и умения на новом практическом материале.

Заключение

В этой работе мы охарактеризовали два ведущих направления комплексного исследования семиотической концептуализации тела и других соматических объектов. Это (1) классификация и изучение признаков соматических объектов (были выделены и охарактеризованы три крупных класса признаков — структурные, физические и функциональные) и (2) идеология, методология и практическое построение компьютерной базы данных «Семиотическая концептуализация тела». В дальнейшем мы надеемся использовать выводы и результаты проведённых исследований для типологического сопоставления семиотических концептуализаций тела в разных языках, этносах и культурах.

Литература

1. *Аркадьев, Крейдлин 2010 — Аркадьев П. М., Крейдлин Г. Е.* Части тела и их функции (по данным русского языка и русского языка тела) // Сборник научных работ к 80-летию Ю. Д. Апресяна. М., 2010 (в печати).
2. *Бирдвистел 1967 — Birdwhistell R. L.* Some body motion elements accompanying sloken American English // O. Thayer (ed.) Communication: concepts and perspectives. Washington D. C.: Spartan, 1967, P. 53–76.
3. *Кадыкова, Крейдлин 2010 — Кадыкова А. Г., Крейдлин Г. Е.* Семиотическая концептуализация тела и его частей. II. Признак «цвет». М., 2010 (в печати)
4. *Крейдлин 2009 — Крейдлин Г. Е.* Тело, эмоции, температура // (отв. ред. проф. С. В. Ионов) Язык и эмоции: номинативные и коммуникативные аспекты. Сборник научных трудов к юбилею В. И. Шаховского. Волгоград: Волгоградское научное издательство, 2009. С. 85–96.
5. *Крейдлин 2010 — Крейдлин Г. Е.* Тело в диалоге: семиотическая концептуализация тела (итоги проекта). часть 1: тело и другие соматические объекты // Компьютерная лингвистика и интеллектуальные технологии (по материалам ежегодной Международной конференции «Диалог» (Бекасово, 26–30 мая 2010 г.)). М., 2010. Вып. 8 (16).
6. *Крейдлин, Переверзева 2009 — Крейдлин Г. Е., Переверзева С. И.* Признак «ориентация части тела» в семиотической картине мира // Молдован А. М. (отв. ред.) «Слово — чистое веселье». Сборник статей в честь А. Б. Пеньковского. М.: Языки славянской культуры, 2009. С. 337–349.
7. *Крейдлин, Переверзева 2010 — Крейдлин Г. Е., Переверзева С. И.* Семиотическая концептуализация тела и его частей. I. Структурные характеристики соматических объектов. М., 2010 (в печати).

Глаголы вращения: лексическая типология[1]

Verbs of rotation: lexical typology

Круглякова В. А. (v.kruglyakova@gmail.com)
Российский государственный гуманитарный университет

Рахилина Е. В. (rakhilina@gmail.com)
Институт русского языка

В работе представлен анализ семантики глаголов вращения на материале 15 языков и опыт составления семантической карты для данного поля. Обсуждаются стратегии склеивания значений и как результат реализации этих стратегий — образование бедных и богатых лексико-семантических систем вращения разных типов.

1. Введение

Лексическая типология становится актуальным и важным лингвистическим направлением и привлекает интерес все большего числа исследователей (см. обзоры Плунгян, Рахилина 2007, Koptjevskaja-Tamm 2008, сборник Goddard (ed.) 2008). Среди наиболее крупных проектов последних лет можно назвать сборник Newman (ed.) 2002 о глаголах позиции, исследование глаголов разрушения Majid et al. 2007, 2008, глаголов движения в воде Майсак, Рахилина (ред.) 2007; сборник о терминах ландшафта (Burenhult, Levinson 2008) и актуальный проект по глаголам боли (Резникова и др. 2008).

Лексическая типология ставит перед собой разнообразные задачи, среди которых изучение того:

- какие значения выражаются лексически;
- какие значения могут лексикализоваться в одной лексеме, а какие выражаются только свободными словосочетаниями;
- лексикализация каких значений является универсальной и наблюдается в большинстве языков, и кодирование каких значений лингвоспецифично и обнаруживается в ограниченном числе языков;
- какие типы значений могут соседствовать в одной лексеме — феномен полисемии;
- какие типы лексических систем образуются в результате совмещения значений.

Данная работа, нацеленная на анализ семантики глаголов описывающих вращение, встраивается в парадигму подобных исследований. Впервые типологический подход к глаголам вращения был применен в работе (Рахилина, Прокофьева 2004), настоящее исследование продолжает и существенно расширяет его. Рассматривается материал 15 языков: русский, польский, сербский, испанский, английский, валлийский, татарский, турецкий, калмыцкий, японский, китайский, алюторский, агульский, хинди, коми. Выделяются семантические параметры, релевантные для этого поля, и предлагается семантическая карта.

В работе применяется метод корпусного исследования: анализируется выборка контекстов, в которых встречаются глаголы вращения и выявляются границы их сочетаемости. Также используются методики полевых исследований.

2. Семантико-синтаксические особенности глаголов вращения

Глаголы вращения — предикаты, описывающие ситуацию, при которой «предмет перемещается по кругу относительно воображаемой точки или движется по кругу относительно воображаемой линии, проходящей через его центр, неоднократно принимая

[1] Исследование поддержано грантом РФФИ 09-06-00364-а

одно и то же положение в пространстве» (НОСС 2004). Эта группа глаголов является удобным объектом лексико-типологических исследований — выражаемое ими значение достаточно простое, оно релевантно для всех культур и присутствуют во всех языках, кроме того, их количество в среднем не превышает 10 единиц, что позволяет провести глубокий анализ.

Глаголы вращения — одна из подгрупп глаголов способа движения (Talmy 1975, 2000). Интересны они тем, что по аргументной структуре отличаются от других групп глаголов движения — для их семантики неважны начальная и конечная точка, направление, способ и среда движения — и совпадают с глаголами позиции. На первый план выходят Trajectory (круг), Trajector (движущийся субъект) и Landmark (ориентир) (ср. Langacker 1988). Очевидно, что наличие до 10 глаголов в рассматриваемой области и их четкая дифференциация не могут основываться исключительно на данных элементах, должны существовать и другие значимые семантические параметры. Таким образом, в задачу представляемого исследования входит, во-первых, выявить параметры, релевантные для лексической зоны вращения, во-вторых, определить, какие из них претендуют на вхождение в универсальный семантический набор, и, в-третьих, определить возможные и запрещенные способы контаминации значений в этой зоне.

3. Лексико-семантические параметры вращения

Данные проанализированных языков показывают, что основополагающим параметром для выбранного нами семантического поля является ось вращения. Противопоставление двух или более глаголов в зависимости от того, вращается ли субъект вкруг собственной оси (карусель, фигурист) или вокруг другого предмета (планеты вокруг Солнца) существует в агульском, алюторском, валлийском, польском, японском языках:

(1) Яп.
Kanransha ga ***mawatte*** *imasu.*
колесо.обозрения NOM вращаться-CNV AUX.PRG-ADR-PRS
‘Колесо обозрения **крутится**’.

(2) Яп.
Kondoru wa wareware no zuzyou o ***senkaishita***
кондор TOP мы GEN над головой ACC вращаться-PST
‘Кондор **кружил** над нашими головами’.

3.1. Зона внешней оси

Внутри каждой зоны выделяются более частные параметры, которые выстраиваются в определенную иерархию. Для зоны внешней оси предлагается:

внешняя ось
рефлексивное не рефлексивное
вихревая траектория круговая траектория
над ориентиром в одной плоскости
с поступательным движением без перемещения

Различается движение по траектории, приближенной к идеальному кругу (электроны, люди в хороводе), и движение по вихревой траектории, т. е. вращение субстанции или сножества мелких однородных предметов, результат которого визуально напоминает воронку (смерч, листья в вихре). При этом маркированным оказывается второй тип вращения, для обозначения которого в японском используется глагол *uzumaku*, в испанском — *arremolinarse*, в калмыцком — *цоонгрх*, в английском — *whirl*.

(3) Норв.
Vind-en har roet seg, men ennå ***virvle-r (*snurrer, *kretser, *roterer)*** *om barnets føtter*
ветер-DEF.SG.M PRAE.PERF успокоиться.PART REFL но все еще кружиться-PRAE вокруг ребенок нога
‘Вихрь утих, но все еще **кружится вихрем** у ног ребенка’.

Вращение вокруг внешней оси по круговой траектории можно, в свою очередь, разделить на вращение в одной плоскости с ориентиром (люди танцуют вокруг костра, лошадь качает воду) и на движение, при котором субъект находится над ориентиром, преимущественно на большом расстоянии от него (ястреб над добычей, муха над тарелкой). Так, в хинди глагол *phirnā* используется для обозначения вращения одушевленных субъектов вокруг внешней оси только в одной плоскости с ними, зона вращения над ориентиром обслуживается другим глаголом, *ma^{n}Drānā*:

(4) *tura^{n}t sabhī pakŠiyo^{n} ne krodhit hokar uske sir par* ***ma^{n}Drānā (*phirnā)*** *aur čīkhnā šūrū kiyā*
тут же все птицы Erg разъяренными став его голове на кружить (Inf) и кричать начали
'Тут же разъяренные птицы стали **кружить** над его головой и кричать'.

(5) *kouve ko pakaRne kī košiš karte hue bheRiyā^{n} peR ke čāro^{n} or* ***phirne (*ma^{n}Drāne)*** *lage*
ворону Acc поймать Gen попытку делая волки дерева Gen четырех сторон кружить(Inf) начали
'Пытаясь поймать ворону, волки стали **кружить** вокруг дерева'.

Вращение вокруг ориентира в одной плоскости с ним в дальнейшем может быть разделено на движение, в процессе которого субъект существенно не перемещается (хоровод) и на вращение, сопровождаемое поступательным движением (падающие листья, танцующие пары). Последний тип хорошо представлен в русском языке — именно его описывает глагол *кружиться*:

(6) *Снег* ***кружится****, летает и тает…*

Завершающим этапом распределения параметров в данной зоне является разделение на вращение, при котором субъект совершает множественные или одинарные обороты (турбина и турникет). По сути, глаголами вращения являются предикаты, описывающие длительное движение со множеством оборотов. Ср. рус. *Фигурист* ***вращается*** *на одной ноге, поднимая вторую и удерживая ее руками* vs. *Мальчик услышал шум и* ***повернулся*** *чтобы посмотреть, что там происходит*. Однако, собранные данные показывают, что ряд языков (калмыцкий, хинди, валлийский и др.) считают это противопоставление нерелевантным — один и тот же глагол одинаково может описывать обе приведенные ситуации:

(7) Исп.
Hab-iendo escuch-ado esas palabras María ***se gir-ó*** *y dio una bofetada a Juan.*
PERF-GER слышать-PART эти слова Мария REFL вращаться-3SG.PST и дать.3SG.PST INDEF пощечина DAT Хуан
'Услышав такие слова, Мария **повернулась** и ударила Хуана'.

(8)
La Tierra ***gir-a*** *sobre su eje.*
DEF Земля вращаться.3SG.PRAE на свой ось
'Земля **вращается** вокруг собственной оси'.

3.2. Зона внутренней оси

Первое, что стоит разделить в этой зоне — вращение при плотном контакте с плоскостью (качение) и вращение, не сопряженное с таковым.

3.2.1. Качение

Во всех рассмотренных языках есть глагол качения, в некоторых языках эта зона делится между двумя глаголами: агульский *ada*$_{ç_0}$*as*, валлийские *ymdreiglo* и *rholio*, китайский *gǔn*, турецкий *yuvarlanmak*, сербские *котрљати се* и *ваљати се* и т. д. Для остальных глаголов вращения в языках эта зона недоступна:

(9) Валл.
Syrthiodd un boncyff coeden gefn y lorri a ***rholio*** *i lawr y llethr*
падать-PRT3Sg one log-tree from back ART lorry AND roll- VN down ART slope
'Упало одно бревно из грузовика и **покатилось** вниз по склону'.

В зоне качения можно выделить вращение в контакте с плоскостью, сопровождаемое поступательным движением (бревно с горы) и вращение без существенного перемещения, при котором субъект совершает многократные разнонаправленные обороты (животные в драке).

В данной оппозиции маркированным является второй член (разнонаправленное вращение), т. к. если в языке есть только один глагол в зоне качения, он покрывает либо исключительно первое значение (в таком случае второе выражается глаголом из другого семантического поля, например, колебания, или является невыразимым в данном языке), как происходит в агульском языке, в коми, либо одновременно оба, что наблюдается в калмыцком, алюторском языке. Не обнаружено такого языка, в котором был бы один глагол качения, распространяющийся только на вторую ситуацию. Ср. валлийские *powlio* и *ymdreiglo*, а также калмыцкий *көлврх* (10), (11).

(10)

Zalu-s bax-d-igə piramid kevtähär og-chk-və, bolәv deerkə-n' un-ad ***kölvr-äd*** *od-və*
Мужчина-PL бревно-PL-ACC пирамида как положить-PRF-PST но верхний-3SG падать-CONV кувыркаться-CONV уходить-PST
'Мужчины сложили бревна в пирамиду, но верхнее упало и **покатилось** по склону'.

(11)

Xoir chonə bazhld-ad hazər deerə ***kölvr-äd*** *noold-və*
Два волк сцепиться-CONV земля на кувыркаться-CONV драться-PST
'Два волка сцепились друг с другом и начали драться, **катаясь** по земле'.

Разграничение вращения с многочисленными vs. одинарными оборотами является конечным и в этой зоне. В некоторых языках (английский, валлийский) глагол, описывающий качение может также расширяться на одинарные неполные обороты (перевернуться с боку на бок). Стоит заметить, что контаминация этих двух значений не является типичной и встречается крайне редко.

3.2.2. Вращение без контакта с плоскостью

Третьей крупной зоной рассматриваемого семантического поля является вращение вокруг внутренней оси без контакта с плоскостью (юла, флюгер). В этой зоне крайне важным оказывается принцип антропоцентричности языка, т. к. одним из ключевых параметров, по которым отличаются квазисинонимичные глаголы вращения, является одушевленность. Так, в хинди глагол *phirnā* описывает упорядоченное круговое движение вокруг собственной оси только неодушевленных предметов[2] (глобус, колесо машины, карусель), а у польского *wiercić się*, напротив, может быть только одушевленный субъект (как правило, это человек):

[2] Если же при глаголе стоит одушевленный субъект, то вращение происходит вокруг внешней оси, вокруг ориентира.

(12)

Niech pan ***się*** *tak nie* ***wierci*** *— powiedzia-ł tata.*
Чтобы пан REFL так не крутиться.3SG.IMV сказать-3SG.PST-M отец
'Не крутитесь так, — сказал отец'.

(13)

Wygrywał ten uczestnik, któremu bąk ***kręcił*** ***się (*wiercił się)*** *najdłużej.*
Выигрывать-3SG.PST тот участник которому юла крутиться.3SG.PST-M REFL дольше всего
'Выигрывал тот участник, у которого юла **крутилась** дольше'.

Для вращения неодушевленных субъектов существенна скорость, высокая и низкая. Интересно, что относительно вращения скорость считается низкой, если человек может различить отдельные обороты, совершаемые субъектом. И наоборот, если субъект вращается настолько быстро, что обороты сливаются в равномерное движение, скорость считается высокой. Этот параметр является релевантным в системах агульского, валлийского, калмыцкого, сербского, татарского языков. Именно по этому параметру противопоставлены английские глаголы *spin* и *revolve*:

(14) *The skater hisses and* ***spins (*revolves)*** *in jump.*
'Фигурист свистит и **крутится** в прыжке'.

(15) *The Boss was rushing from his limo to a* ***revolving (*spinning)*** *door when he turned to answer a fan's shouts.*
'Босс несся от своего лимузина к **вращающейся** двери, но обернулся, чтобы ответить на выкрики фанатов'.

Различение вращения с одинарными и множественными оборотами существенно и в этой зоне семантического поля. В большинстве языков глагол, описывающий вращение вокруг внутренней оси, безразличен к количеству совершаемых оборотов и распространяется на одинарные обороты. Это явление наблюдается в агульском, валлийском, калмыцком, алюторском, сербском, турецком языках, в хинди, коми. Однако есть и противоположные примеры, когда область одинарных оборотов остается недоступной для глагола и для ее описания (как в русском, польском языках) требуется особый предикат. Интересны в этом отношении глаголы вращения коми: от одного корня при помощи суффиксов актантной деривации образуются глаголы, обозначающие одинарные или множественные обороты. Глагол *бергооны* оформлен аспектуальным суффиксом *-оо- (-ал-)*, одним из базовых значений которого является выражение глагольной множественности, поэтому, как и следует ожидать, он обозначает вращение с множественными оборотами (16). При добавлении к этой же основе суффиксов повышающей (*-ед-*) и понижающей (*-ч-*) актантной деривации, получившийся глагол *бергедчыны* обозначает одинарные обороты (17).

(16)

Понм-ыс ***берг-ал-э*** *и мэд-э* *кут-ны* *ассис* *бэж-сэ.*
собака.OBL-POSS3SG крутиться-ITER-PRS.3SG и хотеть-PRS.3SG поймать-INF REFL хвост-ACC.POSS3SG
'Собака **крутится** и хочет поймать свой хвост'.

(17)

Восьтан-ыс ***берг-ед-ч-іс*** *томан пычк-ас.*
ключ-POSS3SG поворачиваться-TR-DETR-PST.3SG замок внутри-ESS/ILL.POSS3SG
'Ключ **повернулся** внутри замка'.

3.3. Дополнительные параметры

Вместе с тем, существуют параметры, которые невозможно поместить в какое-либо конкретное место предлагаемой схемы — они слишком общие и в разных языках могут становиться релевантными в совершенно разных зонах поля. Примером такого параметра является контролируемость. Так, в сербском языке он проявляется в зоне качения, именно на нем основано противопоставление глаголов *котрљати се* и *ваљати се*: первый используется для описания целенаправленного движения одушевленных субъектов (спортсмен специально катится чтобы укрепить мышцы живота), а второй — для нецеленаправленного и неконтролируемого движения одушевленных субъектов (человек поскользнулся и катится с горы). В русском языке этот параметр является релевантным для зоны вращения вокруг внутренней оси (*крутиться* vs. *вертеться*).

4. Стратегии совмещения значений. Семантическая карта

Выделенные параметры и значения не лексикализуются в отдельные глаголы, а совмещаются внутри лексем. Существенным моментом является то, что языки выбирают эти значения не случайным образом, а по определенным правилам. Например, регулярно одной лексемой описываются ситуации вращения в одной плоскости с ориентиром и над ним (наблюдается в валлийском, алюторском, сербском, японском языках). Противоположная ситуация — ни в одном языке не встретилось склеивание «катания» по поверхности и вращения вокруг ориентира. По этому принципу составлена семантическая карта: чем выше вероятность совмещения значений, тем ближе они расположены на карте.

Склеивание нескольких значений в рамках одной лексемы на схеме отражается путем включе-

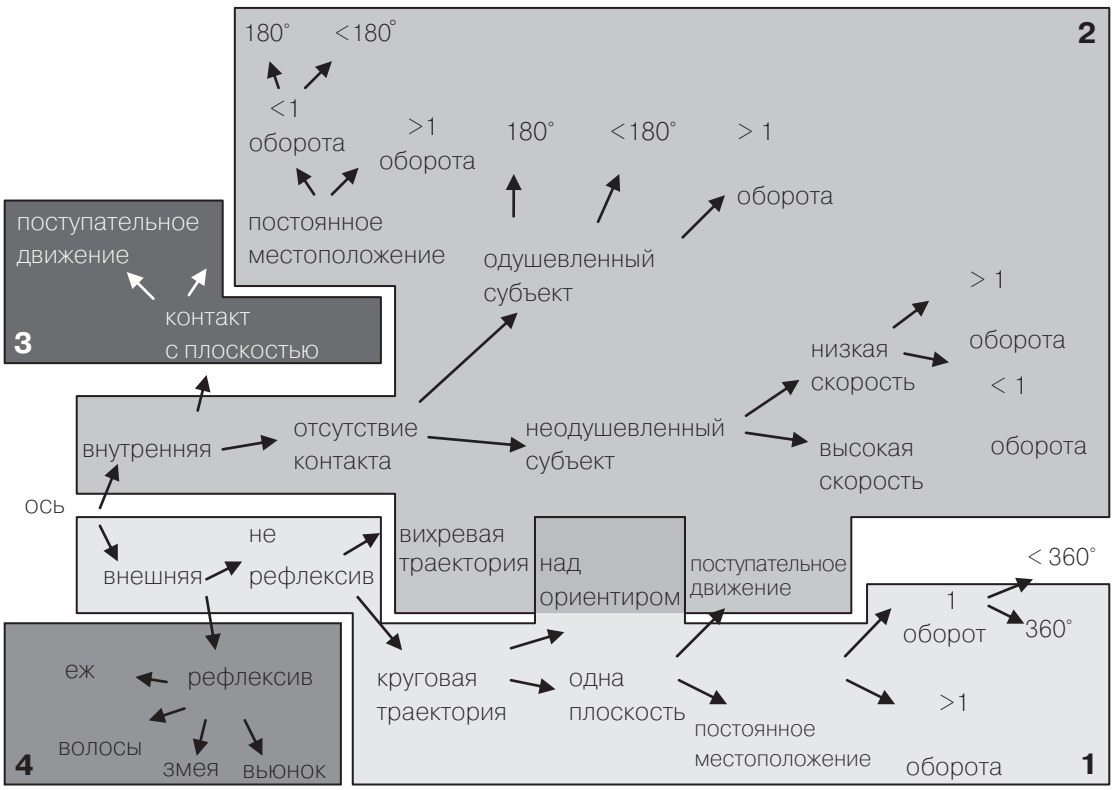


Рис. 1. Система коми-зырянского языка: 1. гэгредлыны, 2. бергооны, 3. тырооны, 4. гартчыны

ния этих значений в одну зону. Число выделенных зон соответствует количеству глаголов в системе определенного языка, а их размер отражает продуктивность глагола. На рис. 1 приведена схема[3] системы глаголов вращения в коми-зырянском языке.

По принципу близости значений было произведено деление семантического поля на 3 базовых зоны: вращение вокруг внешней оси, вокруг внутренней оси без контакта с плоскостью и вращение вокруг внутренней оси в контакте с плоскостью (качение). Это основополагающее разграничение присутствует в подавляющем большинстве языков и является исходной точкой для деления систем глаголов вращения на бедные, средние и богатые. Бедными считаются системы, состоящие из 3 глаголов (минимальное обнаруженное число), как правило, соответствующих выделенным базовым зонам, как это происходит в турецком языке. Средние системы состоят из 4–5 глаголов, в них присутствуют дополнительные лексические оппозиции в ряде зон, это системы агульского, алюторского, китайского, японского языков. Более 6 глаголов формируют богатую систему, для которой количество релевантных семантических параметров в каждой зоне гораздо выше. К этому типу относятся системы глаголов вращения русского, сербского, английского, испанского, валлийского языков. Например, в сербском языке зона вращения вокруг внешней оси делится между тремя глаголами: *кружити*, *окретати се* и *обилазити*. *Кружити* описывает только такое движение, при котором совершаются многократные или неисчисляемые обороты вокруг ориентира, в центре внимания — просто факт кругового движения; при этом акт движения воспринимается как единое целое (лошадь вокруг водоподъемного колеса). Если субъект совершает один или любое конкретное число оборотов, особенно если они преследуют какую-либо цель, то такое потенциально конечное вращение описывается глаголом *обилазити* (жрецы обходят алтарь). *Окретати се* используется для медленного вращения, при котором можно различить отдельные обороты, совершаемые субъектом (хоровод).

[3] Обратим внимание на еще на одну зону — рефлексивное движение (Lindner 1982). Эта область является периферией семантического поля вращения и требует особого обзора, но выходит за рамки данной работы.

5. Заключение

Если заполнить сем карту поля вращения для каждого языка и, наложив их друг на друга, сравнить полученные результаты, то мы получим ответы на вопросы, поставленные в данном типологическом исследовании. В частности, они наглядно показывают, что существуют когнитивно релевантные фреймы вращения, присутствующие в большинстве языков, а системы глаголов вращения, как и все семантическое поле в целом, организованы по определенным правилам.

Исследование проходит в рамках программы, цель которой — доказать, что лексика при всем ее разнообразии, как и грамматика, имеет системный характер: лексикализации подвергаются совершенно определенные значения, входящие в своего рода универсальный набор лексических значений. При этом, как и в грамматических системах, разные значения из этого универсального набора могут объединяться в одной лексеме, следуя стратегии, принятой в данном языке.

Литература

1. *Майсак, Рахилина* (ред.). Глаголы движения в воде: лексическая типология. Москва: Индрик, 2007.
2. *НОСС* — Новый объяснительный словарь синонимов русского языка. 2-е изд. Под рук. Ю. Д. Апресяна. — Москва; Вена: Языки славянской культуры: Венский славистический альманах, 2004 г.
3. *Рахилина Е. В., Прокофьева И. А.* Родственные языки как объект лексической типологии: русские и польские глаголы вращения // Вопросы языкознания, 2004. 1, 60–78.
4. *Резникова Т. И., Бонч-Осмоловская А. А., Рахилина Е. В.* Глаголы боли в свете грамматики конструкций // Научно-техническая информация. Серия 2: Информационные процессы и системы, 2008. 4, 7–15.
5. *Niclas Burenhult, Stephen C. Levinson.* Language and landscape: a cross-linguistic perspective \\ Language Sciences, Volume 30, Issues 2–3, March-May 2008, Pages 135–150
6. *Goddard C. (ed.).* Cross-Linguistic Semantics. Amsterdam: Benjamins, 2008.
7. *Koptjevskaja-Tamm M.* Approaching lexical typology. // Vanhove M. (ed.). From Polysemy to Semantic Change. Amsterdam: Benjamins, 2008.
8. *Langacker.* Concept, image and symbol: The cognitive basis of grammar. B. Mouton de Gruyter, 1991. 149–164.
9. *Lindner S.* What goes up does not necessarily come down. The ins and outs of opposites // Papers from the 18th Regional Meeting, Chicago Linguistic Society. Chicago, 1982.
10. *Bowerman M., van Staden M., Booster J. S.* The semantic categories of "cutting and breaking" events across languages // Cognitive Linguistics, 2007. 18(2). 133–152.
11. *Majid A., Booster J. S., Bowerman M.* The cross-linguistic categorization of everyday events: A study of cutting and breaking // Cognition, 2008. 109(2), 235–250.
12. *Newman J. (ed.).* The Linguistics of Sitting, Standing and Lying. Amsterdam: Benjamins, 2002.
13. *Talmy L.* Semantics and syntax of motion // Syntax and Semantics, vol. 4. N.Y.: Academic Press, 1975. 181–238.
14. *Talmy L.* Toward a cognitive semantics: Vol. II: Typology and process in concept structuring. Cambridge, MA: MIT Press, 2000.

Из каких элементов состоит метаязык лингвистики?[1]

Крылов С. А. (krylov-58@mail.ru)
Институт востоковедения РАН, Институт системного анализа РАН

§ 1. Металингвистические базы данных как компьютерный инструмент инвентаризации элементов метаязыка лингвистики

Вопрос о том, из каких элементов состоит специальный метаязык лингвистики, может решаться разными путями, среди которых — и автоматизированное извлечение устойчивых словосочетаний из корпуса лингвистических текстов[2], и концептуальная систематизация уже извлечённых из текста терминов[3], в том числе в виде формальной онтологии[4], и автоматический анализ соотношений между экспликансами (терминами, входящими в тексты дефиниций) и экспликандумами (понятиями, разъясняемыми с помощью этих дефиниций)[5].

Один из возможных подходов к вышеуказанной проблеме заключается в практическом эксперименте по составлению предметных указателей к книгам и статьям по лингвистике в формате компьютерных баз данных (БД). В данном докладе делается попытка очертить круг содержательных проблем, встающих при попытке описать то, какие содержательные объекты стоят за теми единицами, которые составляют «входы» в предметные указатели[6] к книгам по лингвистике. Проблема словников указателей обсуждается в §§ 2–3, § 5 и в Приложении; проблема упорядочения предметов в указателе — в § 4.

Информационно-технологический продукт, служащий инструментом систематизации элементов метаязыка лингвистики, называется металингвистической базой данных (МБД). Типам МБД посвящён § 2 данного доклада.

§ 2. Содержательные разновидности металингвистических баз данных

МБД, разрабатываемые в интегрированной информационной среде StarLing[7], служат инструментом систематизации знаний о лингвистике (а не напрямую о языке), однако косвенно способствуют также систематизации сведений о языке. Можно выделять две разновидности МБД: (1) метанаучные (МН-) МБД (входы в которые являются металингвистическими проекциями научных текстов по лингвистике) и (2) метаобъектные (МО-) МБД (входы

[1] Данная работа получила поддержку РГНФ (грант № 08-04-00190а).

[2] См., напр., [Митрофанова и Захаров 2009].

[3] См., напр., Никитина 2010.

[4] См.,напр., [Соколова и др. 2008].

[5] См., напр., [Шелов 2009].

[6] «Предметные указатели» понимаются в широком смысле этого слова (то есть как указатели любых элементов специального метаязыка науки (в данном случае лингвистики), представляющих информационный интерес для пользователя, желающего отыскать в тексте информацию об интересующем его предмете. «Предметные указатели» в узком смысле слова — это то же, что терминологические указатели.

[7] Об этой системе см.: [Крылов, Старостин 2006]; [Крылов, Старостин 2007].

в которые являются металингвистическими проекциями языковых сущностей).

Входами в МО-МБД служат, например, характеристики языковых общностей (имена языков и народов, языковых ареалов, эпох бытования языка); нарицательные лингвистические термины (ЛТ); имена языковых единиц (в том числе имена таксономических классов внеязыковых сущностей).

§ 3. Объекты систематизации в металингвистике

Следует прежде всего проводить различие между онтологическим (материальным) уровнем, на котором можно выделить объектное множество (оригинал, универсум) с существующими в нём отношениями и гносеологический (эпистемологический, идеальный) уровень, на котором выделяется модельное множество (модель, теория) с заданными на нём отношениями. Эту модель и строит металингвист, воплощающий её в виде грамматики, словаря, указателя, графа, дерева, атласа и т. п.

Исходные универсумы в лингвистике таковы (см. подробнее Приложение ниже).

I. Универсум языковых явлений
   - IA. Общелингвистический универсум
   - IБ. Частнолингвистический универсум
   - IB. Универсум речевых событий

II. Универсум собственно лингвистики

III. Мир лингвистических моделей

В целом внутри класса III (модельных объектов) выделяются подклассы объектов, прямо корреспондирующие с объектами класса I. Отличие состоит в том, что объекты класса I первичны, т. е. существуют в долингвистическом мире (до и независимо от лингвистов и их деятельности), тогда как объекты класса III (модели) вторичны, т. е. они суть плоды интеллектуальной деятельности лингвистов. Что касается объектов класса II, то они не относятся ни к онтологии, ни к модельным объектам, так как принадлежат миру субъектов познавательной деятельности.

МН-МБД описывают объекты класса II, а МО-МБД — объекты класса III. Однако объекты класса III (гносеологические объекты) являются моделями объектов класса I (онтологических объектов); благодаря этому МО-МБД косвенно способствуют систематизации знаний о языке.

§ 4. Параметры сортировки в металингвистических базах данных

Основные параметры сортировки объектов лингвистических универсумов и их представления в виде организованных систем таковы.

1. Алфавитное упорядочение. Применяется почти ко всем категориям элементарных объектов (кроме дат).

Разновидности: прямое и обратное (обратное применяется обычно к множествам слов или морфологических единиц).

2. Тематическое упорядочение. Основывается на некоторой заранее известной смысловой классификации и иерархизации инвентаризуемых единиц. Ср. тезаурусные словари, генеалогические перечни языков.

3. Хронологическое упорядочение. Возможно для дат появления текстов, дат публикации словарей и грамматик и т. п.; но также и для имён лингвистов при упорядочении их биографий по датам жизни.

4. Квантитативное (статистическое) упорядочение (ср. словарные статьи в частотных словарях). Объектом статистической иерархизации чаще становятся слова, но бывает и собственно металингвистическая иерархизация, напр., имён лингвистов и их работ по степени цитируемости.

Основные способы упорядочения могут осложняться «гнездованием».

5. Речевое (синтагматическое) упорядочение. Линейный порядок фиксации единиц в БД иконически отображает хронологический (линейный) порядок их появления в тексте. Ср. пословные морфологические разборы и т. п.

6. Дерево (в частности, линеаризованное). Ср. деревья синтаксического разбора.

7. Граф (в частности, дерево). Ср. классификационные деревья, семантические графы; таблицы (напр., «парадигм») и т. п.

Классификации могут быть основаны либо на древовидном, либо на универсальном принципе; реально распространена также комбинированная классификация объектов. Оба типа допускают табличное представление. Тематический порядок всегда отражает ту или иную классификацию объектов. Блок-схемы как особый вид классификации представляют собой связные ориентированные графы.

§ 5. Применение металингвистических баз данных для составления словаря семантического метаязыка русских грамматических теорий

Одной из любопытных частных задач, решаемых в русле изложенного подхода к металингвистике, является инвентаризация и систематизация

семантических элементов (далее СЭ), т. е. элементов семантического метаязыка (далее СМ) русских грамматических теорий на материале корпуса специальных текстов, воплощающих эти теории (далее КГТ). Такая работа базируется на интегральном металингвистическом подходе к СМ, разработанном в ряде работ автора (1985–2010 гг.).

В состав СМ включается несколько типов единиц.

1) дефиниционно-семантические элементы (ДСЭ), то есть слова и конструкции того естественного (или «полу-естественного») языка, на котором строятся толкования единиц, встречающиеся в КГТ (нередко толкования в текстах из КГТ заключены в марровские кавычки); разновидностью ДСЭ являются глоссы, используемые при переводе речевых отрезков с незнакомых или малознакомых языков.

2) категориально-семантические единицы (КСЕ), то есть слова и словосочетания, называющие те или иные типы явлений действительности, обозначаемых единицами изучаемого естественного языка-объекта. КСЕ составляют лексикон того СМ, на котором строятся суждения о семантике в КГТ. КСЕ могут быть простыми и составными.

2.1) Примеры простых КСЕ («категориально-семантических элементов», далее КСЭ) — ***ситуация***, ***положение вещей***, ***факт***, ***действие***, ***деятельность***, ***состояние***, ***процесс***, ***качество***, ***количество***, ***признак***, ***свойство***, ***предмет***, ***субстанция***, ***существо***, ***вещество***, ***человек***, ***лицо***, ***деятель***, ***объект***, ***агенс***, ***пациенс***, ***субъект***, ***орудие***, ***способ***, ***место***, ***время***, ***причина***, ***цель***, ***условие***, ***уступка***, ***следствие***, ***мера***, ***параметр***, ***атрибут***, ***характеристика***, ***характер***, ***оценка***, ***множество***, ***элемент***, ***предел***, ***достижение***, ***момент***, ***период***, ***длительность***, ***кратность***, ***интенсивность***, ***мера***, ***сравнение***, ***сопоставление***, ***тождество***, ***бытие***, ***обладание***, ***локализация***, ***пространство***, ***отношение***, ***часть***, ***целое***, ***деталь***, ***пол***, ***возраст***, ***скорость***, ***форма***, ***размер***, ***протяжённость***, ***содержание***, ***отличие***, ***сходство***, ***мужчина***, ***женщина***, ***ребёнок***, ***детёныш***, ***город***, ***река***, ***житель***, ***средство***, ***транспортное средство***, ***страна***, ***небесное тело***, ***явление природы***, ***следование***, ***предшествование***, ***одновременность***, ***фаза***, ***конец***, ***начало***, ***продолжение***, ***отсутствие***, ***наличие***, ***лишение***, ***наделение***, ***владелец***, ***имущество*** и т. п.

2.2) Составные КСЕ («категориально-семантическими комплексов», далее КСК) имеют неэлементарную структуру. В их составе можно выделить стержневой (ключевой, опорный) элемент (см. ниже СЭ, набранные неподчёркнутым **полужирным**) и его модификатор (распространитель, уточнитель) (см. ниже СЭ, набранные подчёркнутыми шрифтами).

Модификаторы бывают двух типов: характеризующий и параметрический.

2.2.1) Характеризующий модификатор (см. ниже ФСЭ, подчёркнутые волнистой линией) выделяет некоторую разновидность или подкласс того, что обозначено стержневым компонентом КСК. Ср. ***действие*** *физическое*, ***действие*** *речевое*, ***действие*** *ментальное*, ***состояние*** *физическое*, ***состояние*** *психическое*, ***состояние*** *устойчивое*, ***состояние*** *временное*, ***процесс*** *предельный*, ***процесс*** *непредельный*, ***предмет*** *считаемый*, ***предмет*** *несчитаемый*, ***предмет*** *одушевлённый*, ***предмет*** *неодушевлённый* и т. п.

2.2.2) Параметрический модификатор (см. ниже СЭ, подчёркнутые пунктиром) позволяет выделить некоторый параметр, атрибут, признак или составную часть того, что обозначено стержневым компонентом КСК. Типичные примеры таких КСК — это: ***действие*** *# орудие;* ***действие*** *# результат;* ***действие*** *# интенсивность;* ***действие*** *# способ; или* ***предмет*** *# размер;* ***предмет*** *# форма;* ***предмет*** *# функция;* ***предмет*** *# устройство; или* ***лицо*** *# пол;* ***лицо*** *# возраст;* ***лицо*** *# национальность;* ***лицо*** *# профессия* и т. п.

В реальных текстах возможно одновременное появление характеризующего модификатора и параметрического: ***предмет*** *считаемый # способ счета (штуками, парами, дюжинами)*.

3) функционально-семантические элементы (ФСЭ) представляют собой модифицирующие компоненты составных лингвистических терминов, обозначающих языковые или речевые единицы. ФСЭ обозначают семантические характеристики языковых и речевых единиц. Они призваны дать семантическую спецификацию тех единиц, которые обозначаются стержневыми компонентами соответствующих терминов. ФСЭ употребляются в составе трёх типов лингвистических терминов — (3.1.) собственно семантических терминов (ССТ); (3.2.) семантически ориентированных терминов (СОТ), не являющихся собственно семантическими; (3.3.) ономасиологических СОТ.

(3.1) В составе ССТ ФСЭ обозначают те или иные семантические характеристики внутренней стороны знаков. Стержневыми компонентами ССТ обычно бывают термины *значение*, *функция*, *смысл*, *семантика*, *употребление*, *использование*.

(3.2) В составе СОТ, не входящих в состав ССТ, ФСЭ обозначают подклассы двусторонних (значимых) единиц языка в целом, выделяемые по признаку наличия тех или иных (соответствующих) семантических характеристик у внутренней стороны этих значимых единиц. Стержневыми компонентами таких СОТ могут быть разные лингвистические термины, но чаще всего ими бывают термины *слова*, *глаголы*, *предикаты*, *имена*, *существительные*, *прилагательные*, *наречия*, *предлоги*, *союзы*, *частицы*, *предложения*, *конструкции*, *словосочетания*, *обороты*, *высказывания*.

(3.3) В составе ономасиологических СОТ (СООТ) ФСЭ используются в качестве заполнителей второй валентности собственно ономасиологиче-

ских терминов. В качестве носителей этой валентности обычно выступают термины *имена*, *названия*, *обозначения*, *наименования*, а также глаголы *означать*, *именовать*, *называть*, *обозначать*, *выражать*, *передавать* и конверсивные по отношению к ним глаголы *именоваться*, *обозначаться*, *выражаться*.

ФСЭ бывают адъективными и субстантивными. ФСЭ в составе большинства СООТ являются субстантивными (исключение составляют такие СООТ, как *реляционные имена*, *одушевлённые имена*, *оценочные обозначения*, *соматические выражения* и т. п.). ФСЭ в составе большинства СОТ, не входящих в состав ССТ, являются адъективными (исключение составляют термины типа *глаголы* ***движения***, *глаголы* ***чувства***, *глаголы* ***речи***, *прилагательные* ***цвета***, *предикаты* ***оценки*** и т. п.).

Субстантивные ФСЭ тождественны соответствующим КСЕ (и потому не нуждаются в отдельной инвентаризации), а адъективные ФСЭ представляют собой адъективные (обычно регулярные, но изредка — супплетивные или квазисупплетивные) дериваты от соответствующих КСЕ. Типичные примеры адъективных супплетивных (или квазисупплетивных) дериватов: ***действие → акциональный***, ***состояние → статальный***, ***отношение → реляционный***, ***восприятие → перцептивный***, ***бытие → экзистенциальный***, ***вещество → субстанциальный***, ***повтор → мультипликативный (итеративный)***, ***событие → эвентуальный*** и т. п.

Приложение. Какие «предметы» могут обозначаться словами, составляющими входы в предметные указатели?

I. Языковые явления

IA. Общелингвистический универсум

(IA.1.) Мир языковой системы: языковая система и ее подсистемы; языковые единицы (ЯЕ); отношения между ЯЕ; члены отношений между ЯЕ (именуемые по их ролям в системе); функции ЯЕ; способы выражения значений; классы ЯЕ; члены классов ЯЕ; языковые структуры; части языковых структур (именуемые по их ролям в структуре); логические связи языковых явлений.

(IA.2.) Речевая динамика: речевые процессы и речевая деятельность (типы, аспекты и компоненты); речевые события, речевые действия (типы, аспекты и компоненты); речевая способность (типы, аспекты и компоненты); речевое варьирование (типы и проявления).

(IA.3.) Языковое функционирование.

(IA.4.) Языковые изменения.

(IA.5.) Языковые сходства и различия.

(IA.6.) Исторические отношения между языковыми общностями.

IБ. Частнолингвистический универсум

(IБ.1.) Исторические языковые общности.

(IБ.1.1.) Языки.

(IБ.1.2.) Надъязыки: генеалогические единства (группы, семьи, макросемьи); контактные единства (языковые союзы); ареальные единства (территориальные множества языков).

(IБ.1.3.) Подъязыки (варианты одного языка): диалекты (территориальные, профессиональные, сословные); исторические этапы языка; жанровые подъязыки.

IБ.2. Ареалы распространения языков: континенты, регионы, страны, провинции; населённые пункты.

(IБ.3.) Частнолингвистические единицы.

(IБ.3.1.) Знаковые единицы.

(IБ.3.1.1.) Лексические единицы: слова и их варианты (семантические и формальные); фразеологизмы разной протяжённости и их варианты (семантические и формальные), паремии (пословицы, поговорки, загадки, крылатые слова и выражения, ходячие афоризмы).

(IБ.3.1.2.) Морфологические единицы.

(IБ.3.1.2.1.) Знаменательные морфологические единицы (основы; корни).

(IБ.3.1.2.2.) Грамматические (служебные) морфологические единицы (форманты, аффиксы, флексии и их эквиваленты).

(IБ.3.1.3.) Конструкции (синтаксические и лексико-синтаксические).

(IБ.3.1.4.) Иероглифы (цифры и другие идеографические символы).

(IБ.3.2.) Фигуры выражения (фонемы, варианты фонем, буквы, слоги).

(IБ.3.3.) Фигуры содержания (семы, понятийные категории, индивидные сущности [8] и т. п.).

IB. Мир речевых событий

(IB1.) Словесность (множество зафиксированных текстов на некотором языке).

(IB1.1.) Специальная устная словесность (фольклор): памятники фольклора; сказители; регионы распространения традиции; даты фиксации произведения; персонажи.

(IB1.2.) Специальные памятники письменности (филологический мир): памятники письменности; внешние субъекты

[8] Индивидные сущности соотносимы с объектами действительности (инвентарь которых фиксируется, в частности, в энциклопедиях и других источниках знаний о мире).

коммуникации (авторы; адресаты; переписчики; переводчики; комментаторы; рецензенты и критики); даты (создания; публикации; переиздания; перевода).

(IB1.3.) Мир неспециальных устных и письменных связных текстов (дискурсов) и их связных частей (высказываний): тексты, высказывания, участники речевых актов (говорящий, слушающий, аудитория), а также разнообразные характеристики связных тестов и высказываний.

IB2. Вхождения речевых знаков-экземпляров (tokens) (ВЭ).

(IB2.1.) ВЭ двусторонних знаков;

(IB2.2.) ВЭ мыслей-представлений или референтов.

(IB2.3.) ВЭ звучаний-представлений (произнесений и написаний), т. е. сигналов

II. Универсум собственно лингвистики

(II.1) лингвисты (в том числе лингвисты-непрофессионалы);

(II.2) лингвистические школы и направления;

(II.3.) лингвистические кружки, общества, ассоциации и т. п.;

(II.4) места, где протекает деятельность лингвистов (континенты, страны, провинции, города) [9]

(II.5) учреждения, где протекает деятельность лингвистов (научные: академии, институты, отделы, секторы и т. п.; учебные: вузы, школы; кафедры; ведомства; редакции; фирмы и т. п.).

(II.6.) Лингвистические события.

(II.6.1.). Памятники лингвистической и прелингвистической (филологической, фольклорной, диалектографической) деятельности (ПЛД): грамматики, словари, издания памятников, словарные картотеки и т. п.; даты (создания; публикации; переизданий; переводов) самих ПЛД.

(II.6.2.). Цитирование этих ПЛД (ссылки на ПЛД; отклики, содержащие ссылку на данный ПЛД; даты публикации откликов; авторы откликов).

III. Мир лингвистических моделей

III.1. Описания языков (словари, грамматики и т. п.)

III.2. Описания текстов и речевых отрезков: хрестоматии текстов, издания памятников, продукты транскрипции и транслитерации, переводы, фонетические сонограммы, комментарии, глоссы, формальные представления в виде морфологических и синтаксических «разборов», синтаксических графов (в частности, деревьев зависимостей и составляющих), цепочки трансформационного вывода, толкования примеров и т. п.

III.3. Описания ЯЕ: словарные статьи, правила (законы), исключения к ним и т. п.

[9] «Места» можно подразделить на кабинетные (учреждения) vs. полевые (соотносимые с объектами из универсума IБ.2, см. выше).

Литература

1. *Митрофанова О. А., Захаров В. П.* Автоматизированный анализ терминологии в русскоязычном корпусе текстов по корпусной лингвистике // Кибрик А. Е. (ред.), Компьютерная лингвистика и интеллектуальные технологии. Вып. 8 (15). По материалам международной конференции «Диалог'2009» (Бекасово, 27–31 мая 2009 г.). М.: РГГУ, 2009, с. 321–328 (http://www.dialog-21.ru/dialog2009/materials//49.htm).
2. *Шелов С. Д.* Мир лингвистической теории сквозь призму ее терминологии (к итогам лингво-компьютерного анализа одного лингвистического текста) // Исследование познавательных процессов в языке (Серия: Когнитивные исследования языка. Выпуск V). М., 2009.
3. *Никитина С. Е.* Семантический анализ языка науки. На материале лингвистики. М.: Либроком, 2010. — 146 с.
4. *Соколова Е. Г., Кононенко И. С., Загорулько Ю. А.* Опыт систематизации знаний и интернет-ресурсов для Портала знаний по компьютерной лингвистике // Кибрик А. Е. (ред.), Компьютерная лингвистика и интеллектуальные технологии. Вып. 8 (15). По материалам международной конференции «Диалог'2009» (Бекасово, 27–31 мая 2009 г.). М.: РГГУ, 2009, с. 465–470 (http://www.dialog-21.ru/dialog2009/materials/html/72.htm)
5. *Крылов С. А., Старостин С. А.* Интегрированная информационная среда StarLing и её использование в сфере корпусной лингвистики // Лауфер Н. И., Нариньяни А. С., Селегей В. П. (ред.), Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог'2006» (Бекасово, 31 мая — 4 июня 2006 г.), М.: Издательство РГГУ, 2006. С. 303–307 (http://www.dialog-21.ru/dialog2006/materials/html/Krylov.htm).
6. *Крылов С. А., Старостин С. А.* Создание и переработка лексических баз данных в интегрированной информационной среде StarLing // Кобозева И. М., Нариньяни А. С., Селегей В. П. (ред.), Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог'2007» (Моск. обл., 1–6 июня 2007 г.), М., Наука, 2007 (http://www.dialog-21.ru/dialog2007/materials/html/49.htm.

Особенности лексико-морфологического анализа при извлечении информационных объектов и связей из текстов естественного языка

Peculiarity of lexical-grammatical analysis for object extraction from natural language texts

Кузнецов И. П. (igor-kuz@mtu-net.ru),
Сомин Н. В. (somin@post.ru)

Институт проблем информатики РАН, Москва

Анализируется опыт построения семантико-ориентированных лингвистических процессоров, выделяющих структуры знаний из текстов естественного языка (ЕЯ). Одним из важнейших компонент таких систем является блок лексико-морфологического анализа. В процессе развития данного класса процессоров этот блок постоянно совершенствовался и приобрел много новых функций, выходящих за рамки возможностей существующих блоков подобного типа. Данный блок генерирует лексические, морфологические, семантические признаки слов, выявляет простейшие формы естественного языка, имеет специальные средства настройки на предметную область и на особенности текстов ЕЯ. В работе рассматриваются эти функции.

1. Введение

1.1. Системы с возможностью обучения языку

На протяжении последних 20 лет в ИПИ РАН активно развивается область, связанная с анализом текстов естественного языка (ЕЯ) с целью извлечения полезной информации, формирования структур знаний и их использования для решения прикладных задач — поисковых, логико-аналитических. В 90-х годах был создан класс экспериментальных систем, которые обладали уникальной особенностью — их можно было обучать естественному языку и в дальнейшем использовать для задач фактографического поиска и принятия решений. Это системы ДИЕС [3, 17], СПРУТ, LOG [5], ЭССЕИСТ [6], ИКС[3, 4]. Эти системы создавались в рамках соответствующих пионерных проектов ИПИ РАН, поддерживаемых госбюджетом. Для обучения языку был разработан специальный интерфейс, с помощью которого вводились не только морфологические признаки, но и семантическая компонента каждого нового слова [17]. Для существительных нужно было указать, к какому классу они относятся. Для глаголов — модель управления и т. д. Для ввода новых слов не требовалось каких-либо специальных знаний. В системе ДИЕС это мог делать практически любой грамотный носитель языка — русского (для ИКС — и английского). В результате формировалась база лингвистических знаний слов и словосочетаний (с их семантикой), которая использовалась для лексико-морфологического и синтаксического анализа предложений ЕЯ с их отображением на структуры знаний, составляющие базу предметных знаний (БЗ). Система осуществляла полный разбор предложений с извлечением структур знаний (фактов). При этом учитывались случаи полисемии глаголов, восстанавливалась информация, заданная в неявном виде, и многое другое. Фактографический поиск и другие приложения осуществлялись на уровне структур знаний.

Для создания подобных систем требовались специальные языки представления знаний и инструментальные средства их обработки. Язык — это структурный объект на всех его уровнях, — от поверхностного до семантического. Для обработки

конструкций языка были созданы: язык расширенных семантических сетей (РСС), обеспечивающий представление текстов ЕЯ на уровне структур знаний с любой требуемой точностью, и язык ДЕКЛ для преобразования структур в виде РСС [1, 2, 17].

В тоже время, развитие подобных систем требовало достаточно трудоемкой работы по вводу новых слов. Системы не могли работать с предложениями, где было много незнакомых для них слов. В связи с этим системы типа ДИЕС нашли применение для создания языков экспертных систем, содержащих ограниченное количество слов и форм, достаточных для ввода экспертных знаний. Одна из них — это система СПРУТ, предназначенная для выявления организованных преступных групп

1.2. Объектно-ориентированные процессоры.

В связи со сказанным в конце 90-х годов в рамках проектов ИПИ РАН начало развиваться другое направление, при котором не требовалось отображения семантики всех предложений на структуры знаний в БЗ [8, 9, 10]. Учитывался тот факт, что определенные категории пользователей интересуются конкретной информацией, которая встречается в текстах ЕЯ. Нужно извлекать из текстов только эту информацию. Данное направление возникло в связи с прикладными разработками для ГУВД г. Москвы. Их проблемы заключались в наличии потоков документов на ЕЯ (сводок происшествий, справок по уголовным делам, обвинительных заключений и др.), в которых было много полезной информации. Это фигуранты, их адреса, телефоны, оружие, автотранспорт и др. Следователей и аналитиков интересовали именно такого сорта объекты и связи между ними. Использование типовых БД требовало громадной работы для их заполнения.

В связи с этим в ИПИ РАН была инициирована работа по созданию лингвистических процессоров (ЛП), обеспечивающих автоматическое выделение их текстов ЕЯ информационных объектов и связей с формированием структур знаний в БЗ. Такие ЛП были названы ***объектно-ориентированными*** (в некоторых работах — ***семантико-ориентированными***). Для успешного создания подобных ЛП у разработчиков уже имелась достаточная база. В результате была создана система «Криминал», обеспечивающая автоматическое извлечение структур знаний из текстов ЕЯ и их использование для решения логико-аналитических задач — для следователей и аналитиков. В данной системе не требовалось вводить морфологические и другие характеристики слов. Для этого был создан блок ***лексико-морфологического анализа*** (ЛМА), который анализирует текст и строит семантическую сеть (РСС), названной ***пространственной структурой текста*** (ПС-текста). Последняя обрабатывается блоком ***синтактико-семантического анализа*** (ССА), который (на языке ДЕКЛ) анализирует ПС-текста и формирует на РСС структуру, представляющую объекты и связи между ними. Такие структуры образуют БЗ.

Отметим, что блок ЛМА написан на языке Си++, при использовании которого на определенных этапах формализации текстов возникают существенные трудности. В тоже время, чем больше функций берет на себя блок ЛМА, тем в большей степени снимает трудности дальнейшего процесса формализации, который осуществляется блоком ССА [13, 14, 20].

Поэтому в последующих проектах ИПИ РАН (АНАЛИТИК, ПОТОК и др.) блок ЛМА постоянно усовершенствовался [7, 18]. В процессе выполнения проектов объектно-ориентированный ЛП использовался в различных предметных областях для формализации различных корпусов текстов. В связи с этим в блок ЛМА постоянно вводились новые возможности. В данной статье обобщается опыт разработчиков по построению объектно-ориентированных ЛП. Статья посвящена описанию особенностей блока ЛМА, выполняющего сложные функции по анализу текстов ЕЯ и обеспечивающего необходимой информацией блок ССА для данного типа ЛП.

2. Особенности объектно-ориентированных ЛП

Наш опыт показывает, что при наличии потока документов, требующих обработки, учесть все формы и особенности ЕЯ, используемые при описании многих объектов и связей, и построить сколь либо полную «модель языка» — неразрешимая задача. Поэтому требуется постоянное совершенствование ЛП.

В связи с этим рамках проектов ИПИ РАН развивается направление, когда программа объектно-ориентированного ЛП отделяется от ***лингвистических знаний*** (ЛЗ). Последние определяют всю процедуру анализа (см. ниже). ЛЗ имеют вид декларативных структур, которые легко менять и настраивать. В нашем случае роль таких структур выполняют фрагменты РСС [12–14]. Настройка ЛП осуществляется только за счет разработки ЛЗ.

Задача ЛП — поддерживать ЛЗ. При использовании подобных ЛП облегчается настройка на корпуса текстов, особенности предметной области. Корректировать ЛЗ может человек, обученный формализму РСС и знакомый с элементами математической лингвистики. Ему не нужно уметь программировать. Тогда возникает вариант, когда один человек может настраивать лингвистический процессор — находить ошибки и устранять их.

Как говорилось ранее, объектно-ориентированный ЛП состоит из блоков лексико-морфологического и синтактико-семантического анализа.

Блок **лексико-морфологического анализа** (ЛМА), выделяет из документа слова и предложения и выдает в виде семантической сети (ПС-документа), представляющей последовательность компонент (слов в нормальной форме, чисел, знаков) и их основные признаки. Блок ЛМА имеет три основных подсистемы:

- Лексический анализатор, который ответственен за правильное деление входного текстового потока на абзацы, предложения и слова (формирует лексические признаки слов);
- Морфологический анализатор, осуществляющий морфологический анализ всех слов текста (приводит слова в нормальную форму и формирует для них морфологические признаки);
- Систему предметных словарей, призванную распознать в тексте характерные термины (формирует семантические признаки).

Блок ЛМА имеет свои лингвистические знания (ЛЗ) — средства ***параметрической настройки***, позволяющие учитывать разнообразие текстовой типологии, и набор ***предметных словарей*** (словарь стран, регионов России, имен, профессий и др.) для придания словам и словосочетаниям дополнительных семантических признаков [13, 18].

Блок **синтактико-семантического анализа** (ССА) путем анализа ПС-документа выделяет объекты и связи. На их основе строит другую семантическую сеть, представляющую ***семантическую структуру документа*** (СС-документа), называемую также ***содержательным портретом*** [12, 15, 19]. Этот блок включает в себя ***базу лингвистических знаний*** (ЛЗ), которая содержит правила анализа текста во внутреннем представлении (РСС). Они определяют работу ЛП [13–19].

Блок ССА управляется ЛЗ, за счёт которых обеспечивается:

- извлечение информационных объектов (лиц, организаций, событий, их места);
- выявление связей объектов; например, связей лиц с организациями, адресами и др.;
- анализ глагольных форм, причастных и деепричастных оборотов с выявлением фактов участия объектов в тех или иных действиях;
- идентификация объектов с учетом анафорических ссылок и сокращенных наименований;
- выявление связей действий с их местом или временем (где и когда имело данное действие или событие);
- анализ причинно-следственных и временных связей между действиями и событиями.

Особенности блока ССА описаны во многих статьях, в том числе трудах конференций Диалог [8–16]. Гораздо меньше внимания уделялось описанию работы блока ЛМА. В данной статье будет восполнен этот пробел.

Блок ЛМА [7], разработанный для русского и английского языков, основан на традиционной для таких блоков схеме словарей. Однако, помимо этого, в блоке ЛМА присутствует еще ***словарь обобщенных основ***, позволяющая обрабатывать и новые слова, см. п. 5.

Блок ЛМА приводит слова в нормальную форму и присваивает им признаки, которые делятся на три группы: — лексические признаки (слово с большой буквы, большими буквами, с точкой на конце или это отдельная буква и др.)

- морфологические признаки (грамматическая категория слова, число для существительных и т. д.);
- семантические признаки (имя, организация, оружие и др., а также ключевые слова, относящиеся к соответствующему типу объектов).

Предусмотренный лексикографический анализ обеспечивает автоматическое деление текста на самостоятельные части (например, выделение документов из сводок) и определение начала и конца предложения, а также начала и конца абзаца.

Выходная информация бока ЛМА (т. е. ПС-текста) сохраняет порядок предложений в тексте, разделяя их фрагментами типа SENT, и порядок слов в предложении. При этом каждое слово представляется с его признаками, см. п. 7.

3. Предметные области и тексты

В настоящее время имеется большой опыт использования объектно-ориентированных ЛП в различных предметных областях, где требуется выделение различных объектов из корпусов текстов со своими особенностями. В данном разделе мы постараемся обобщить эти особенности и связанные с ними трудности, которые требовали постоянного совершенствования блока ЛМА. С какими предметными областями и текстами мы имели дело:

3.1. Документы криминальной милиции

Работа делалась по заказу ГУВД г. Москвы [9]. Была создана система «Криминал», в БЗ которой были введены: сводки происшествий по (более 500 тыс. происшествий), справки по уголовным делам (несколько сотен), обвинительные заключения (около сотни), записные книжки фигурантов (около сотни). Система обеспечивает выделение фигурантов, их примет, связей, организаций, дат, документов, номеров счетов, оружия (всего до 40 типов объектов) с указанием характера их участия в криминальных действиях .

3.2. Резюме (для приема на работу) на русском и английском языках

Работа инициировалась компанией HEADHUNTER и имела целью автоматическую обработку архивов произвольно написанных резюме и их представление в формате сайта данной компании [16]. Была создана система, выделяющая из резюме атрибуты человека, места его работы, учебы, соответствующие периоды времени, знание языков и т. д. Система отлаживалась в несколько этапов. Вначале на выборках в различных областях (информационные технологии, банковское дело, финансы, юриспруденция и др.) по 200 резюме. Далее отладка шла на специально подобранных «критичных» резюме (которые при обработке давали шумы по тем или иным типам объектов), из которых составлялись специальные выборки. Система работала на сайте упомянутой компании, чтобы автоматически переводить резюме пользователей, поступающих через Интернет, в формат сайта.

3.3. Документы о терроризме на русском языке.

Работа носила инициативный характер с целью внедрения в крупный проект. Система дополнительно выделяла руководящих лиц, правительственные организации, террористов (как свойство фигурантов), террористические организации, орудия преступления, время и место событий и т. д., а также связи и участие лиц в тех или иных действиях. На первом этапе она отлаживалась на массиве в 300 документов, относящихся к террористической деятельности. В дальнейшем отладка шла на материалах СМИ, в том числе, взятых из Интернет [18]. Была также разработана ДЕМО-версия для обработки документов на английском языке.

3.4. Документы о памятниках культуры

Работа носила инициативный характер — делалась для Министерства культуры. Система выделяет из текстов тип памятника (скульптура, монумент), кто является автором, создателем, время, место и многое другое.

Во всех случаях (за счет средств настройки блоков ЛМА и ССА) удавалось добиться требуемого качества работы ЛП [10, 13, 17].

Отметим высокое разнообразие перечисленных предметных областей, которое определяется не только в различием выделяемых объектов и связей. Еще большие отличия можно наблюдать в «стиле» текстовых сообщений, связанных с предметными областями. В понятие «***стиль***» мы включаем весь комплекс особенностей, присущих определенной группе текстов. Сюда входят:

- лексика предметной области, включая всю совокупность специфических терминов предметной области;
- коммуникативный тип текста: художественное произведение, техническая или аналитическая статья, новостное сообщение, приказ, PR- текст (например — реклама);
- структурный тип текста: связный текст, список, таблица, математическая формула;
- инструмент создания текста (имеется в виду текстовый редактор или генератор текста, с помощью которого получен текст);
- способ грамматического оформления текста, под которым понимается следование стандартным правилам орфографии языка (проставление необходимых знаков препинания и разделителей, позволяющих структурировать текст);
- следование принятой в языке орфографии, что выражается в количестве орфографических ошибок или нарочитого введения искаженной лексики.

Отметим, что резкое увеличение разнообразия текстовой типологии, с которой мы столкнулись в различных предметных областях. В значительной степени это вызвано бурным распространением Интернет и тем фактом, что порождение текстов все в большей мере стали осуществлять люди различной степени подготовки и грамотности. Как следствие — наличие значительного количества специальных разделителей, отсутствие знаков препинания, большое количество сокращений, ошибок и многое другое. Отсюда следуют дополнительные требования к компонентам блока ЛМА и средствам их настройки. Рассмотрим их подробнее.

4. Лексический анализатор

Эта компонента блока ЛМА имеет дело с целым рядом взаимосвязанных задач, решение которых совершенно необходимо для успешной работы всего ЛП. Рассмотрим их особенности.

Прежде всего, решается **задача структуризации текста**. Дело в том, что текст в современной информационной среде — сложно структурированный объект. И его структура должна быть распознана и аккуратно передана блоку ССА, поскольку от правильного распознавания структуры текста в значительной степени зависит корректность всего анализа. Поэтому задача структуризации распадается на цепочку локальных задач.

4.1. Задача выделения лексем

При выделение из входного потока лексем (слов, знаков препинания, разного рода раздели-

телей и др.) имеют место следующие трудности. Современный деловой текст содержит большое количество лексем, являющихся техническими, административными и фирменными названиями, телефонами, шифрами, номерами автомобилей, адресами электронной почты и Интернет и проч., содержащими цифры, буквы и разделители практически в произвольной комбинации. Такие знаки, как «-», «.» и «,», доставляют много хлопот при их анализе, в одних случаях являясь разделителями лексем, а в других — нет.

4.2. Задача выделения предложений

В виду огромного разнообразия текстовых «стилей», по отношению к современным текстам становится трудно говорить о предложении. Скорее следует говорить о «сильносвязанных» отрезках текста, в которых идет речь об одном объекте или одной ситуации, в которой участвуют несколько взаимодействующих объектов. В результате само понятие «предложение» резко расширяется, включая в себя, помимо обычных предложений (с точкой в конце), еще массу различных текстовых отрывков: ячеек таблицы, элементов списка и прочих, грамматическое оформление которых нетрадиционно.

4.3. Задача выделения абзацев

Абзацем мы называем отрезок текста из одного или нескольких предложений, связанных единой темой. Опять-таки расплывчатость этого определения позволяет трактовать его широко. Однако для блока ССА понятие абзаца является весьма важным, поскольку многие его механизмы направлены именно на идентификацию и совмещение объектов внутри одной темы. Лексический анализатор содержит в своем составе ряд алгоритмов, выделяющих абзацы, причем — разных типов.

Надо сказать, что задачи выделения предложений и абзацев весьма нетривиальны. Трудности выделения абзацев главным образом связаны с тем, что хорошо различимые разделители абзаца — пустые строки, отступы, границы клеток таблицы — теряются или искажаются при преобразовании текстов. Но гораздо большие трудности возникают при идентификации предложений. Дело в том, что современные пользователи Интернет вообще не считают необходимым ставить точки в конце предложения. В тоже время точка активно используется в качестве ограничителя сокращений, разделителя между частями несловарного компонента (электронного адреса, многозначного числа, банковского номера и пр.). Кроме того, разделителем предложения может являться не только точка, но и другие знаки («;», «:», «!», «?», «|» и т. д.). В результате задача разбиения текста на предложения становится просто головоломной шарадой, требующей учета массы разного рода частных правил и исключений.

4.4. Задачи унификации текста

Естественный язык — система необычайно многовариантная: один и тот же смысл (приблизительно) может быть выражен умопомрачительным количеством его текстовых выражений. Задача лексического анализатора: унифицировать написание отдельных слов и сокращений, привести к стандартной форме написание ряда стандартных словосочетаний. Трудность тут в том, чтобы выявить наиболее употребительные лексемы и словосочетания, требующие унификации.

К этой сложности примыкает проблема обнаружения и (по возможности) исправления ***опечаток и грамматических ошибок***. В современных текстах их — громадное количество, и бороться с ними — задача из сложнейших. Кроме того, в современных текстах, особенно из Интернет, намечается тенденция нарочитого переделывания и перевирания слов, типа «*ацкий ужос*» или «*падстол*». Начинает формироваться целая интернетная «феня». В связи с этим потребуется постоянная корректировка языковых словарей и правил составления предложений.

Еще одна важная функция лексического анализатора — определение ***лексических признаков*** слов. Примеры такого рода признаков: «Слово из кириллицы с прописной буквы». «слово из кириллицы из прописных букв», «разделитель», «слово из латинских букв» и проч., всего — около 20 лексических признаков. Лексические типы являются важной дополнительной информацией, облегчающей работу как морфологического анализатора, так и блока ССА.

Наконец, лексический анализатор для ряда слов способен выполнить семантический анализ, определяя по формальному виду слове его ***семантическую категорию***. К этому случаю относятся, скажем, сокращения имен и отчеств: прописная буква, за которой идет «.». Например «*А.», «Н.», «J.».* Еще примеры идентифицируемых семантических классов: «адрес электронной почты», «Интернет-адрес» (URL), «целое число», «число с дробной частью». Собственно, определение семантического класса каждого слова или словосочетания является одной из задач всего ЛП. И чем раньше такой класс будет определен, тем легче дальнейший анализ.

5. Морфологический анализатор

Задача морфологического анализатора — нормализация слов, определение морфологических признаков лексем, а также (в ряде случаев) — на-

хождение его семантического класса. Отметим, что к настоящему времени разработан целый ряд морфологических анализаторов русского языка, среди которых упомянем лишь некоторые: [21, 22, 23, 24].

5.1. Базовая схема анализа

Первоначально была реализована базовая схема анализа [5]. Считается, что каждое слово имеет постоянную часть (основу) и переменную часть. Последняя образует словоизменительную парадигму или класс окончаний. Были накоплены два словаря: словарь классов окончаний (СКО), в котором хранятся все возможные парадигмы русского языка и словарь основ (СО), в котором хранятся основы слов со ссылками на соответствующий класс окончаний.

Например, слово «*бытие*» имеет основу «*быти*» и класс окончаний за номером 1759, содержащий окончания в именительном, родительном, дательном, винительном, творительном и предложном падежах, а именно: «*е», «я», «ю», «е», «ем», «и*» (множественного числа это слово не имеет). Соответственно в СО имеется запись «*быти 1759*», а в СКО под номером 1759 закодирована парадигма с указанными окончаниями.

Отметим, что в общем случае в СО может быть несколько записей с одинаковой основой (но с разными классами окончаний), а на один и тот же класс окончаний может ссылаться несколько слов с разными основами. Возможны случаи пустой основы (пример: «*хорошо*»-«*лучше*») и пустого класса окончаний (для неизменяемых слов). Кроме основы и вариантов окончаний, в СКО хранятся морфологические признаки, соответствующие определенному классу окончаний в целом (постоянная морфологическая информация) и каждому окончанию парадигмы в отдельности (переменная морфологическая информация). Так, для класса 1759 в качестве постоянной информации хранятся признаки существительного, среднего рода, неодушевленности и второго склонения, а для каждого окончания хранится признак соответствующего падежа.

Алгоритм морфологического анализа при наличии данных словарей сводится к следующему. Для слова рассматриваются все варианты его разбиения на основу и окончание. Если для данного варианта разбиения находится основа, а в соответствующем ей классе окончаний находится вариант окончания, то данный морфологический разбор является корректным и слово получает морфологические признаки, взятые из постоянной и переменной частей морфологической информации. В общем случае может быть найдено и выдано несколько вариантов морфологического разбора, что известно, как морфологическая омонимия.

5.2. Морфологический анализ незнакомых слов

В принципе предложенная схема анализа вполне корректна. Однако на практике ее успешное использование достаточно проблематично. Дело в том, что такая схема предполагает ручную разработку обоих словарей. И заметим — не только первоначальную разработку, но и их постоянное пополнение. Последнее обстоятельство особенно неприятно: в русском языке — более 100 тыс. слов общеупотребительного назначения и миллионы специальных терминов. Кроме того, перестройка вылилась в активную языковую экспансию: в русскоязычных текстах стало использоваться огромное количество англоязычных слов, которые никогда не входили в классические словари русского языка. В результате при обработке таких текстов система «натыкалась» на множество слов, отсутствующих в СО. Фактически требовалось ежедневное пополнение словаря. Но в то же время, создание таких словарей требует высокой лингвистической квалификации и исключительной тщательности. Составление достаточно полных морфологических словарей — кропотливая работа, требующая десятилетий.

Выход из описанной ситуации известен — обработка незнакомых системе слов «по аналогии» [24, 25]. В нашей реализации этого метода использовался третий словарь — «***словарь хвостов основ***» (СХО). В словарь записываются все 1-буквенные, 2-буквенные, 3-буквенные и т. д. «хвосты» основ (первые буквы основ отбрасываются) с указанием соответствующего класса окончаний. Было решено, что в СХО не будет одинаковых «хвостов», а его класс окончаний вычисляется из статистических соображений — по максимуму основ в СО, имеющих данный «хвост» и данный класс окончаний. Если слово не находится в словаре СО, то та же схема анализа повторяется, но уже с помощью пары словарей СХО-СКО.

В реализации словари СО и СХО были слиты в один словарь, за которым закрепилось название обобщенного словаря основ (ОСО), в результате чего все варианты анализа, — как точные, так и по аналогии, — выявляются за один проход по словарю. Кроме того, был разработан способ сжатия словарной информации, который позволил хранить все словари в оперативной памяти существующих на момент реализации (1996 г.) компьютеров (объем словаря на 90 тыс. основ составляет 894 КБ).

5.3. Борьба с морфологической омонимией

Ясно, что использование обобщенного словаря основ ОСО может приводить к лишним вариантам морфологического анализа. Было предложено два достаточно эффективных способа борьбы с морфологической омонимией.

Первый способ — эмпирический алгоритм, отбрасывающий наименее вероятные варианты морфологического анализа. Такая «зачистка» вариантов выполняется по многим критериям, учитывающим наличие слова в СО, длину основы с СХО, часть речи. Кроме того, эмпирический алгоритм расставляет все варианты разбора в порядке их вероятности. Такое ранжирование необходимо для ряда приложений, когда используется только один вариант морфологического анализа.

Второй способ — частичный синтаксический анализ. Дело в том, что в предложении слово вступает в синтаксические связи с другими словами, и выявление этих связей позволяет отбросить варианты морфологического анализа, этим связям не удовлетворяющих. Прежде всего было реализовано распознавание двух конструкций: полного согласования и генетической цепочки.

5.4. Особенности блока английской морфологии

Помимо русского морфологического словаря был создан и английский морфологический словарь. Он использует уже разработанное для русской морфологии программное обеспечение, которое оказалось возможным адаптировать к специфике английского языка. Блоки английской и русской морфологии выдают практически одни и те же морфологические характеристики. Это позволяет использовать для синтаксико-семантического анализа англоязычных текстов те же средства, что и для русского языка. В результате появилась возможность записывать лингвистические знания для этих языков в одном и том же формализме.

Общий объем словаря основ блока английской морфологии — около 85 тыс. Тем не менее, для повышения качества работы этого блока в него был введен ряд специфических для английского языка алгоритмов, которые в основном касаются отсева лишних вариантов морфологического анализа. Дело в том, что слова английского языка чрезвычайно омонимичны. Очень часто одно и то же слово может быть и существительным, и глаголом, и прилагательным. В блоке английской морфологии реализованы алгоритмы, позволяющие в ряде случаев корректно отбрасывать лишние варианты (другие варианты отсеиваются в процессе синтактико-семантического анализа). Блок ЛМА был модифицирован для работы с предметными словарями английского языка, которые удалось совместить со словарями русского языка.

6. Система предметных словарей

Предметные словари (стран, имен собственных, организаций, профессий, видов оружия и др.) состоят из терминов. Множество словарей образует систему.

Система предметных словарей (СПС) предназначена для распознавания в тексте слов и словосочетаний, специфичных для конкретной предметной области. Им присваиваются признаки принадлежности к определенной семантической категории. Будем называть этот процесс идентификацией терминов словаря. Такая принадлежность является основой выделения объекта. В предметном словаре может быть или термин, представляющий объект определенного типа (но таких объектов может быть достаточно много), или характеристическое слово, опираясь на которое можно начинать распознавание объекта — на уровне синтактико-семантического анализа.

Видимо, без СПС не обходится ни один серьезный проект лингвистического процессора. В нашей разработке СПС встроена в блок ЛМА. Причина этого — главным образом в быстродействии. Поиск в СПС предполагает частые обращения к ней, а потому требуется высокая эффективность поиска, чего трудно достичь без использования универсальных языков программирования. В нашем случае программное обеспечение СПС написано на Си++.

Структурно СПС состоит из произвольного количества ***словарей***, являющих определенный семантический класс. В каждом из словарей может содержаться произвольное количество ***словарных записей***. Под записью в тривиальном случае понимается термин (однословный или многословный). Однако простыми терминами словарные объекты не ограничиваются — там могут содержаться ***словарные шаблоны***, описывающие группу терминов. Возможности описания словарных шаблонов будут приведены ниже. В настоящее время разработаны более 20 предметных словарей; среди них: «Улицы г. Москвы», «террористические организации», «оружие», «известные личности».

6.1. Требования к предметным словарям

К СПС, помимо эффективности, предъявляются еще ряд требований, важнейшими из которых являются:

1) *Требование множественности*. Информационная система может иметь несколько предметных словарей различного содержания, причем число словарей заранее не ограничивается и может динамически пополняться. Общее число словарей может превышать несколько сотен.
2) *Требование к объему предметного словаря.* Каждый словарь, разумеется, содержит множество записей, список которых может расширяться. В нашей постановке задачи предполагается, что объем словаря не может быть заранее ограничен каким-либо фиксированным числом, а должен быть потенциально неограниченным и определяться только

вычислительными мощностями и объемом оперативной памяти и накопителей.

3) *Требование к подготовке информации*. Подготовка текстового материала для загрузки словаря, естественно, выполняется специалистами в соответствующих предметных областях. Поэтому форма исходного вида словаря должна быть максимально простой.
4) *Требование вариативности поиска*. Должна быть предусмотрена корректная обработка случаев, когда написание термина в тексте так или иначе не соответствует каноническому виду термина в словаре.

Если первые три требования носят в основном технический характер, то удовлетворить требованию вариативности поиска в условиях естественного языка — задача весьма непростая. Дело не только в том, что термин может стоять в любом из падежей, что не дает возможности напрямую совместить текст с предметным словарем (эта проблема решается с помощью морфологического анализатора). Основная трудность в другом: в социуме имеет хождение множество вариантов употребления одного и того же термина, и указать их все для разработчика предметного словаря является непосильной задачей. Вот примеры.

Как правило, названия улиц записаны в именительном падеже. Например, «*проживает по адресу Б. Академическая ул. д. 6–18*». Иногда встречается дательный падеж: «*по Б. Академической*». Гораздо более усложняет дело вариативность сокращений и перестановки слов. Например, канонический вид названия одной из улиц Москвы — «*Щипковский 1-й пер.*». Однако, встречаются в текстах написания: «*1-й Щипковский пер.*», «*1-ый Щипковский переулок*», «*п-к 1-вый Щипковский*» и другие варианты. Отметим, что возможна не только перестановка и вариативное написание слов, но и выпадение или добавление слов. Например, «*Туполева Академика наб.*» может быть названа как «*набережная Туполева*», а «*Тихий туп.*» иногда добавляют пояснение «*ул. Тихий туп.*». Кроме того, некоторые сокращения, применяемые авторами текстов, далеко не однозначны. Например «*С.*» может означать «*Северный*» или «*Старый*»; «*Б.*» может означать «*Большой*», а может быть сокращением имени, например «*ул. Б. Галушкина*».

6.2. Возможности предметных словарей

Подключение новых словарей может значительно усилить ЛП в плане выделения объектов. Однако для того чтобы словари в самом деле стали действенным и удобным механизмом, необходимо, чтобы они обладали рядом нетривиальных возможностей.

В нашей версии СПС реализованы несколько таких возможностей.

Во-первых, идентификация термина в любом числе и падеже. Например, если в словаре есть термин «*программный продукт*», то в тексте будут распознаваться и соответствующим образом идентифицироваться термины «*программного продукта*», «*программных продуктов*» и т. д. Распознавание выполняет программное обеспечение системы предметных словарей, использующее блок морфологического анализа.

Во-вторых, допускается несколько вариантов написания одного и того же термина. Дело в том, что средствах СМИ и многих других текстах пользуются различными вариантами именования одного и того же объекта, в том числе сокращенным описанием. Например, если в тексте встретилось *Путин, Меркель, президент Франции* и т. д., то понятно, о ком идет речь. Для приведения таких словосочетаний к стандартному виду в словари введена специальная запись. Например, в словаре ФИО может иметь место запись:

Меркель Ангела
= Ангела Меркель
= А. Меркель
= Меркель

В данном примере основной термин — «*Меркель Ангела*». К нему будут приводиться все остальные написания этого имени, записанные после символа «=». Эта возможность особенно эффективна при выявлении не только ФИО известных деятелей, но и названий организаций (включая их сокращения), географических названий и др. При этом блок ССА осуществляет дополнительную фильтрацию, например, когда в тексте несколько лиц с фамилией *Меркель* или рядом со словом *Меркель* стоит какое-либо имя, не представленное в предметном словаре.

В-третьих, в предметные словари введена возможность описания группы терминов, у которых лишь первое слово фиксировано, а остальные могут быть описаны с помощью совокупности признаков (лексических и морфологических). Реализованы, так называемые, ***словарные шаблоны***. Например, в словаре допустима строка:

заведующий {NOUN,КЕМ}

Такая запись в словаре профессий означает, что подходящими под этот шаблон терминами могут быть все словосочетания, начинающиеся со слова «*заведующий*», за которым идёт существительное (NOUN) в творительном падеже (КЕМ): «*заведующий складом*», «*заведующий библиотеками*» и т. д. Кроме того, в качестве шаблона можно употреблять имя другого (или того же самого) словаря. Это дает возможность точнее указывать те варианты, которые допускает шаблон. Фактически на словари возлагаются элементы синтаксического анализа, позволяющие значительно уменьшить количество записей в словаре, а также облегчить работу блока ССА.

В-четвертых, имеется возможность управлять лексическим и морфологическим анализами в процессе распознавания терминов словарей. Так, например, в словаре террористических организаций может быть указано:

Организация эта
= ЭТА\!

Это означает, что, благодаря признаку «\», слово «*эта*» в процессе идентификации морфологическому анализу не подвергается (т. е. его каноническая форма совпадает с написанием). И, кроме того, благодаря признаку «!» идентификация совершается, если в тексте слово «*ЭТА*» записано прописными буквами. Эти возможности позволяют повысить точность распознавания, отсеивая ложные вхождения.

Отметим, что язык записи терминов в словарях чрезвычайно прост. Термин пишется в своей канонической форме на отдельной строке (включая, разумеется, указанные выше. дополнительные возможности). Поэтому ввод новых терминов или даже создание новых словарей может быть выполнено пользователем или оператором-лингвистом, не знакомым с особенностями работы ЛП.

Помимо указанных возможностей имеется еще ряд специальных операторов настройки, позволяющих управлять идентификацией терминов для тех или иных словарей.

7. Пространственные структуры

Текст ЕЯ — это сложный структурный объект, который в процессе его формализации проходит множество уровней преобразования. На первом уровне работает блок ЛМА, который формирует РСС, называемую ***пространственной структурой текста*** (ПС-текста). Далее следуют преобразования, осуществляемые блоком ССА, которые приводят к формированию ***семантической структуры*** (СС-текста) для БЗ.

Рассмотрим особенности ПС-текста. Информация об абзацах и предложениях представляется в виде фрагмента SENT, с помощью которого представляется:

- позиция первого слова предложения относительно начала входного потока;
- признак начала абзаца и количество разделительных строк;
- номер строки, на которой расположено первое слово предложения.

Для каждого слова (и для каждого варианта его разбора) блок выдает фрагменты типа LR, задающих последовательность слов. В каждом из фрагментов представлено:: нормализованное слово и его порядковый номер. Далее следуют его признаки. Вот некоторые из них: NAME0 — слово начинается с прописной буквы, HEAD_ — слово полностью состоит из прописных букв, NAME1 — инициалы, POINT — пункт, HEAD_1 — слово с прописной буквой, NUM) — целое число, NUM_F — число с дробной частью, ENGL — слово из букв латинского алфавита, WEB_C — URL (адрес Интернет), MAIL_E — адрес электронной почты, FIRST_ — признак первого слова на новой строке, LETT — слово из одной буквы и т. д. (морфологические и семантические признаки).

Фрагменты типа LR и SENT вместе с выделенными признаками — это семантическая сеть (РСС), которая в дальнейшем проходит множество уровней преобразования, осуществляемое блоком ССА.

В общем случае блок ЛМА выдает несколько вариантов разбора. Эта ситуация является весьма типичной. Например, слово «*стекло*» является и существительным и глаголом. Тогда в ПС-текста, помимо фрагмента LR для первого варианта разбора, генерируются фрагменты LD (с их признаками) для других вариантов. Отсев вариантов осуществляется блоком ССА в процессе обработки ПС-текста и построения семантической структуры [12, 14].

8. Средства параметрической настройки

Опираясь на опыт построения ЛП для различных предметных областей (см. п. 3), чтобы постоянно учитывать все новые особенности текстовой типологии, в блок ЛМА были введены средства управления лексико-морфологическим анализом, названые средствами ***параметрической настройки***. Эти средства относятся к ЛЗ и размещаются в отдельном файле. Они имеют вид списков, оформленных в виде фрагментов РСС со своими именами. Имена играют роль операторов и определяют вид анализа.

Всего реализовано 27 типов фрагментов. Из них 18 относится к блоку лексического анализа, 5 — блоку морфологического анализа и 4 — предметным словарям.

Лексическое оформление текста — один из самых вариативных аспектов, сильно меняющихся от задачи к задаче. В связи с этим аппарат лексической настройки потребовал значительного развития в плане разработки новых операторов (заданных в виде фрагментов). Рассмотрим их, разделив операторы на смысловые группы.

8.1. Средства идентификации начала и конца предложения

- Если слово, указанное во фрагменте NEW_ SENT, записано в тексте с прописной буквы и находится в начале строки, то оно рассматривается как начало нового предложения.

- Если в тексте встречается одно из слов (символов, знаков), указанных во фрагменте END_SENT, то оно считается концом предложения.
- Фрагмент ABBR задает список сокращений с точками на конце, которые считаются цельными словами и точки не рассматриваются как конец предложения
- Фрагмент SEPARATOR задает символы, которые всегда являются разделителями слов.

8.2. Средства для замены или удаления некорректных символов или слов

- Фрагменты LETTER_CH и WORD_BAD задают замены (или удаление) нежелательных слов или знаков в тексте.
- Фрагменты BEG_SYMB задают набор удаляемых знаков в начале слова, а END_SYMB — в конце.

8.3. Средства унификации и синонимичных замен

- Фрагмент SYNON задает список синонимичных слов, которые заменяются на слово из первой позиции.
- Фрагмент TERMIN_ заменяет слова, записанные на второй и последующих позициях, на слово в первой позиции.
- Фрагмент SIGN_MANY задает повторяющиеся символы, следующие один за другим (например, набор черточек) на один символ (черточку).

8.4. Средства настройки морфологического анализатора.

- Фрагмент MORF определяет генерацию морфологических признаков слова в виде фрагментов ПС-текста.
- Фрагмент NOMO задает список слов, для которых устанавливается запрет на нормализацию и морфологический анализ.
- Фрагмент NOMOE задает для слов дополнительные признаки, которые вставляются в ПС-текста.

Это необходимый набор операторов, без которых (как оказалось) трудно обеспечить качественный лексико-морфологический анализ многих текстов ЕЯ, и следовательно, качественную работу всего объектно-ориентированного ЛП.

Заключение

В данной статье рассмотрены особенности блока лексико-морфологического анализа, используемого в объектно-ориентированных лингвистических процессорах (ЛП) при формализации текстов ЕЯ, т. е. для извлечения из них информационных объектов, признаков и связей. Блок обладает уникальными возможностями, с помощью которых обеспечивается устойчивая и качественная работа ЛП при обработке массивов документов на ЕЯ в различных предметных областях: «Криминалистика», «Резюме», «Терроризм», «Памятники культуры» и др.

Литература

1. *Кузнецов И. П.* Семантические представления// М.: Наука. 1986 г. 290 с.
2. *Кузнецов И. П., Шарнин М. М.* Продукционный язык программирования ДЕКЛ. Сб. Система обработки декларативных структур знаний Деклар-2// ИПИ РАН, 1988.
3. *Кузнецов И. П., Шарнин М. М.* Интеллектуальный редактор знаний на основе расширенных семантических сетей// Сб. Системы и средства информатики. Вып. 5. М. Наука, 1993.
4. *Кузнецов И. П.* Гипертекстовые технологии на семантической основе// Сб. Системы и средства информатики. Вып. 7. М. Наука, 1995.
5. *Любушкина Л. А., Михеев А. С., Соловьева Н. С., Сомин Н. В., Фрейдлин И. Я.* LOG — программа, ведущая диалог на естественном языке // Вторая всесоюзная конференция по ИИ «ВКИИ-90». Минск: Центрпрограммсистем, 1990.
6. *Карунин А. Б., Соловьева Н. С., Сомин Н. В.* ЭССЕИСТ — программа, ведущая диалог с базой знаний на естественном языке. // В кн.: Социальная информатика-93 / Сб. научн. трудов под ред. Колина К. К. и Суслакова Б. А. — М., 1993. — С. 168–174
7. *Сомин Н. В., Соловьева Н. С., Шарнин М. М.* Система морфологического анализа: опыт эксплуатации и модификации// Системы и средства информатики, Вып. 15, 2005, стр. 20–30.
8. *Кузнецов В. П.* Автоматическое выявление из документов значимой информации с помощью шаблонных слов и контекста// Труды межд. Семинара Диалог 98? Т. 2. Казань: ООО «Хетер» 1998.
9. *Кузнецов И. П.* Методы обработки сводок с выделением особенностей фигурантов и происшествий// Труды межд. Семинара Диалог 99. Т. 2. Тарусса, 1999.
10. *Кузнецов И. П., Кузнецов В. П., Мацкевич А. Г.* Система выявления из документов значимой информации на основе лингвистических знаний в форме семантических сетей // Труды межд. Семинара Диалог 2000. Т. 2. Протвино, 2000.
11. *Кузнецов И. П.* Лингвистический процессор для автоматического выявления из текстов значимой информации с ее компоновкой в рамках указанных шаблонов// Труды международного семинара Диалог-2001. Том 2. Протвино, Наука, 2001.
12. *Kuznetsov I., Matskevich A.* System for Extracting Semantic Information from Natural Language Text // Труды международного семинара Диалог-2002 по компьютерной лингвистике и ее приложениям. Т. 2. Протвино, Наука, 2002.
13. *Кузнецов И. П., Мацкевич А. Г.* Особенности организации базы предметных и лингвистических знаний в системе АНАЛИТИК // Труды конференции Диалог-2003. Протвино, 11–16.06 2003, стр. 373–378.
14. *Kuznetsov I., Kozerenko E.* The system for extracting semantic information from natural language texts// Proceeding of International Conference on Machine Learning. MLMTA-03, Las Vegas US, 23–26 June 2003, p. 75–80.
15. *Кузнецов И. П., Мацкевич А. Г.* Англоязычная версия системы автоматического выявления значимой информации из текстов естественного языка// Труды международной конференции «Диалог 2005», Звенигород, 2005.
16. *Кузнецов И. П., Мацкевич А. Г.* Семантико-ориентированный лингвистический процессор для автоматической формализации автобиографических данных// Труды международной конференции «Диалог 2006», Бекасово, 2006, с. 317–322.
17. *Кузнецов И. П., Мацкевич А. Г.* Семантико-ориентированные системы на основе баз знаний (монография)// М.: МТУСИ, 2007 г.,173с.
18. *Кузнецов И. П. Сомин Н. В.* Англо-русская система извлечения знаний из потоков информации в среде Интернет // Сб. ИПИ РАН, 2007 г.
19. *Кузнецов И. П.* Объектно-ориентированная система, основанная на знаниях в виде XML-представлений// Сб. ИПИ РАН, Вып. 18. 2008 г., с. 96–118
20. *Kuznetsov I. P., Kozerenko E. B.* Linguistic Processor “Semantix” for Knowledge extraction from natural texts in Russia and English// Proceeding of International Conference on Machine Learning, ISAT-2008. 14–18 July,
21. *Segalovich I.* A fast morphological algorithm with unknown word guessing included by a dictionary for a web search engine // MLMTA-2003. http://download.yandex.ru/company/iseg-las-vegas.pdf
22. *А Коваленко.* Вероятностный морфологический анализатор русского и украинского языков// http://www.keva.ru/stemka/stemka.html
23. *Сокирко А. В.* Морфологические модули на сайте www.aot.ru. Диалог-2004. Верхневолжский, 2–7 июня 2004 г. http://www.aot.ru/docs/sokirko/Dialog2004.htm
24. *Сегалович И., Маслов М.* Русский морфологический анализ и синтез с генерацией моделей словоизменения для не описанных в словаре слов // Диалог'98, Казань, ООО «Хэтер», 1998. http://download.yandex.ru/company/DLG98-MM2.pdf

Прилагательные в составе номинаций лица[1]

Adjectives and personal nouns

Кустова Г. И. (galinak03@gmail.com)

Московский педагогический государственный университет

В работе рассматриваются механизмы включения прилагательных в номинацию — на материале сочетаний прилагательных с личными существительными, ср. *болтливая продавщица, честный исполнитель* и под.

1. Материал, цели и задачи

Естественными номинациями лица являются именные группы, содержащие существительные и / или местоимения: ***Петя*** *устал*; *Звонил* ***брат***; ***Проводник*** *билеты проверяет*; ***Эта продавщица*** *сведет нас с ума* и т. д. В диалогическом режиме такие номинации с трудом включают характеристики, выраженные прилагательными. Есть всего несколько основных случаев, когда это происходит: относительные прилагательные (такие сочетания практически образуют цельную номинацию): *дежурный офицер*; относительные и качественные прилагательные в функции рестриктивного (ограничительного, выделительного) определения (они обозначают признак, по которому данный предмет выделяется из множества): *Подай мне* ***шестую коробку*** */ Возьми (вон то)* ***красное платье***. В «родовых» высказываниях прилагательное представляет свернутую предикацию: ***Опытный инженер*** *не сделает такую ошибку в расчетах* (‘Если инженер опытный, то…’). В большинстве же случаев прилагательные попадают в предикативную позицию (*Сидоров —* ***опытный инженер***).

Более широко прилагательные включаются в номинации в нарративном (повествовательном) режиме (о нарративе см. [Падучева 1996]; [Шмид 2004]). Но и здесь это включение связано с определенными условиями и происходит по определенным правилам. Далее речь пойдет преимущественно о режиме повествования (когда автор стремится изобразить, наглядно представить ситуацию — это не обязательно художественный текст, может быть и газетный), но будут также привлекаться примеры, относящиеся к режиму сообщения (мы не пользуемся термином «диалогический режим», поскольку речь идет о письменных текстах, для нас важно лишь различать «синхронную» точку отсчета, когда наблюдатель находится внутри интервала ситуации и сообщает актуальные характеристики, обнаруживаемые на этом интервале, и ретроспективную точку отсчета, когда автор высказывания сообщает о факте в прошлом и приписывает характеристики на основе знания или обобщения).

Будут рассмотрены типы АТРИБУТИВНЫХ ВКЛЮЧЕНИЙ В НОМИНАЦИИ.

Прежде всего, номинации с прилагательными используются в художественных (а нередко и газетных) текстах для идентификации персонажа (референции). Так, в романе М.А. Булгакова «Мастер и Маргарита» Коровьев может именоваться как *клетчатый тип*, Азазелло (в сцене на бульваре) несколько раз назван *неизвестный гражданин*. Такие номинации — особый художественный прием: они обозначают персонажа не с точки зрения повествователя, а с точки зрения другого персонажа (Берлиоза, Бездомного, Маргариты) и тем самым обеспечивают механизм смены точки зрения в тексте. Однако здесь прилагательное существенной роли не играет, т. к. такая номинация должна рассматриваться как цельная (в этом смысле *клетчатый тип* и *клетчатый*, *неизвестный гражданин* и *незнакомец* не отличаются).

Далее мы будем рассматривать случаи, когда прилагательное выполняет самостоятельную роль в номинации и является отдельным элементом сообщения. Основной вопрос, который будет нас интересовать, — как прилагательное взаимодейству-

[1] Работа выполнена при поддержке РГНФ, проект № 08-04-00183а. Все примеры извлечены из Национального корпуса русского языка. Из соображений экономии места примеры даются в сокращении.

ет с другими элементами структуры предложения: в первую очередь, с определяемым существительным (далее — просто существительное), а также с предикатом (или другими элементами группы предиката).

Поскольку есть множество семантических типов как существительных (*яблоко* — натуральный класс, *еда* — функциональное обозначение), так и прилагательных (постоянный физический признак (*рыжий*, *маленький*); постоянный нефизический признак (*умный*, *честный*, *добрый*); поведение (*болтливый*, *предприимчивый*, *непоседливый*, *вежливый*, *упрямый*); оценка (*противный*, *нахальный*) и др.), мы вынуждены ограничить материал следующими классами слов.

Будут рассматриваться только личные существительные (названия лиц), а из них — только такие, которые включают внутренний предикат:

ФУНКЦИОНАЛЬНЫЕ: *сварщик, стоматолог, прокурор, юрист, секретарь, директор, царь, пастух, вратарь* и др.;

РЕЛЯЦИОННЫЕ: *брат, отец, вдова, друг, враг, сосед, однокурсник, однофамилец, соотечественник, иностранец, незнакомец, любимец* и др.;

РЕЗУЛЬТАТИВНЫЕ: *автор, создатель, убийца, оскорбитель, победитель, угонщик, нарушитель, похититель, беглец, дезертир, эмигрант, приезжий, потерпевший* и др.;

АКТУАЛЬНЫЕ: *всадник, пловец, лыжник, водитель, пассажир, зритель*, *слушатель, сосед* (‘сидящий рядом’), *участник*, *партнер*, *свидетель*, *собеседник* и др. (см. [Арутюнова 1976]; [Шмелев 2002]).

Класс реляционных существительных разнороден, среди них есть термины родства (*брат, отец*) и другие почти «не-глагольные» номинации (*иностранец*, *однофамилец*). Мы будем брать только существительные с очевидными и «полноценными» встроенными предикатами (*друг* → ‘дружить’, *враг* → ‘враждовать’, *сосед* → ‘жить по соседству’, *незнакомец* → ‘не знаком’).

Не рассматриваются:

качественные имена (*болтун, вольнодумец, грязнуля, жадина, знаток, игрок, крикун, либерал, любитель, мечтатель, остряк, трус, хвастун, хитрец* и под.) и собственно оценочные имена (*дурак, подлец, негодяй*): они тяготеют к предикативной позиции и в номинациях (т. е. для идентификации) используются редко и в специальных условиях ([Арутюнова 1976: 349]; [Шмелев 2002: 220]);

демографические номинации (*девушка*, *старик*), т. к. они не содержат внутреннего предиката.

Что касается имен собственных, они будут привлекаться в качестве фона.

Из прилагательных мы будем рассматривать постоянные нефизические признаки, поведение и оценку, которые в нашем случае образуют один большой класс, т. к. они являются обобщением поведения, действий человека в различных ситуациях и могут обозначать как постоянную характеристику (*болтливая соседка*), так и ее актуальное «поведенческое» проявление на конкретном временном интервале (ср.: *Еле отделались от болтливой продавщицы*).

Есть два основных механизма включения прилагательного в номинацию:

а) ДЕНОТАТИВНОЕ ПРИСОЕДИНЕНИЕ (т. е. присоединение к денотату): прилагательное относится непосредственно к денотату существительного и характеризует его как бы «минуя» существительное, вне связи с тем, как назван денотат существительным: *болтливая продавщица*, *болтливая медсестра*, *болтливая соседка* — это характеристика поведения денотата независимо от типа номинации (болтливость не связана с реализацией функций продавщицы или медсестры и не характеризует их);

б) ПРЕДИКАТНОЕ ПРИСОЕДИНЕНИЕ: *бездумный исполнитель* — прилагательное связано с предикатом, который заключен в самом существительном (в данном случае это просто транспозиция наречно-глагольного сочетания *бездумно исполняет*, т. е. это характеристика не вообще человека, а характеристика исполнения им функции).

Мы отдаем себе отчет, что термин «предикатное присоединение» не слишком удачен, но надеемся, что контексте данной работы и в оговоренном смысле он будет пониматься однозначно.

В рамках этих двух механизмов есть разновидности, о которых будет сказано в соответствующих разделах.

2. Денотативное присоединение

2.1. Свободное денотативное присоединение

Общая функция денотативного присоединения, независимо от типа существительного, состоит в информационном насыщении текста — сообщении дополнительной информации. При **свободном** денотативном присоединении прилагательное не связано семантически ни с внутренним предикатом существительного, ни с предикатом предложения.

2.1.1. С функциональными существительными:

После очередного похода в магазин девица уже ощутимо скрипела зубами, особенно когда ***заботливая продавщица*** *сообщила ей, что «вот эта пара имеется с розовыми шнурками».* [Наши дети: Подростки (2004)]; *А вот в киоске «Мороженое» на Ухтомской улице* ***доброжелательная пожилая продавщица*** *никак не могла понять, зачем такая книга нужна, — уже давно никто ни на что не жалу-*

ется. [«Столица», 1997.08.26]; ***Улыбчивые продавщицы**, к которым я обращался с вопросом об «олд малэйжиа мьюзик», предлагали в ответ эстраду 60-х*. [«Домовой», 2002.09.04]; *В подземном переходе на «Библиотеке имени Ленина» выбрала гвоздики диковинного желто-коричневого цвета и попросила у **безразличной продавщицы** хотя бы кусок газеты — укутать беззащитную природу*. [Ольга Новикова. Женский роман (1993)]; ***Равнодушная продавщица**, облачённая в несвежий халат, сидела за двухколесной синей тележкой*. [Андрей Волос. Недвижимость (2000)]; *Как ты думаешь, перегнули вы немного с разоружением-то? — серьезно спросил Федор **неразговорчивого грузчика***. [Н. А. Островский. Как закалялась сталь (1930–1934)].

Прилагательное обозначает денотативно присоединяемую характеристику лица (в приведенных примерах, т. е. в режиме повествования, это поведение лица на выделенном временном интервале), которая просто увеличивает объем сообщаемой в предложении информации. Поэтому, если заменить функциональное существительное на другую номинацию, ничего не изменится, ср.: *Я попрощалась с **болтливой консьержкой** и выскочила на улицу* и *Я попрощалась с **болтливой старушкой***.

В режиме сообщения денотативно присоединяемые прилагательные выражают постоянные поведенческие характеристики: *Процедуры избиений на этом окончились, так же как и посещения **вежливого юриста*** [Вадим Кожевников. Щит и меч. (1968)]; *Иногда с ними* [приезжал] ***напористый юрист**, и тогда их трое*. [Владимир Маканин. Человек свиты (1988)]; *Похоже, в столице начиналась большая политическая игра и **жизнелюбивый прокурор** стал мячиком для разогрева команд перед главным турниром*. [Виктор Мясников. Водка (2000)].

2.1.2. С реляционными существительными:

*Люди Четвертого Века уже купили в Москве жилье, поэтому они свирепо конфликтуют со всеми, разрушают мосты и обижают **безропотных соседей*** [Улья_Нова. Инка (2004)]; *Мне очень жаль, что я так ничего и не узнала об этом **загадочном друге** В. Пятницкого*. [«Лебедь» (Бостон), 2003.11.09]; *А теперь BAR перестала быть частью грандиозного плана двух **предприимчивых друзей** на пути к личному успеху*. [«Формула», 2002.02.15]; *«Носи, носи! — усмехается **суровый незнакомец**. — Я тебе и шляпку выдумаю»*. [А. Аверченко].

2.1.3. С результативными существительными:

*Человек со странным именем Август не спрашивает, кто послал электрограмму с приглашением на выставку византийской миниатюры или откуда взялся **элегантный убийца** трех палачей, заботливо доставивший его к порогу вебстеровского дома*. [Александр Пятигорский. Древний Человек в Городе // «Октябрь», 2001]; *Он* [сорт] *принадлежит народу, нужен ему, независимо от судьбы его **легкомысленного автора***... [Владимир Дудинцев. Белые одежды (1987)].

2.1.4. С актуальными существительными:

Беззаботные пассажиры (*забота была моя, забота меня дурманила*) *гордились властителем-машинистом, покуривали, обменивались знающими улыбками, ложились, дремали*. [В. В. Набоков. Другие берега (1954)]; *Убедившись в том, что Иван один, и прислушавшись, **таинственный посетитель** осмелел и вошёл в комнату*. [М. А. Булгаков. Мастер и Маргарита (1929–1940)].

2.2. Обусловленное денотативное присоединение

Наряду со свободным существует **обусловленное** присоединение, когда прилагательное оказывается в сфере действия предиката данного предложения или других элементов предложения и семантически связано с этими элементами, обусловлено ими. При этом обычно возникают какие-то дополнительные семантические отношения (причинные, условные):

Фельдмаршал [Кутузов] *не счел нужным посвятить **болтливого губернатора*** [Ростопчина] *в свои планы относительно древней столицы* [А. Архангельский. Александр I (2000)] — ‘не посвятил из-за болтливости, опасаясь разглашения’; — *А как это Кравцов согласился? — усмехнулся мэр, хотя на самом деле вполне догадывался, кто мог так повлиять на **трусоватого прокурора** города Приволжска*. [Сергей Таранов. Черт за спиной (2001)] — ‘повлиял настолько, что прокурор преодолел свою трусоватость’; *И до сих пор я вспоминаю с теплым чувством о той самоотверженной работе, которую **этот поистине благородный юрист** вынес на своих плечах в защиту русского народа*. [Г. А. Соломон (Исецкий). Среди красных вождей (1930)] — ‘защищать народ — благородно’; *Он тотчас же рассказал: некий **наивный юрист** представил Столыпину записку, в которой доказывалось, что аграрным движением руководили богатые мужики, что это была война «кулаков» с помещиками*. [Максим Горький. Жизнь Клима Самгина. (1928–1935)] — ‘с его стороны наивно так думать’; *Пока джентльмен пытается освободиться от **назойливых незнакомцев**, девушка исчезает за ближайшим углом* — ‘хочет освободиться, т. к. они назойливы, докучают’; *Грохот. **Испуганные пассажиры** спрашивают: — Что случилось?* [Коллекция анекдотов: железная дорога (1970–2000)] — ‘испугались из-за грохота’; *А в антракте **обиженные зрители** устроили директору скандал*. [Юрий Никулин. Жизнь на колесах (1979)] — ‘скандал в результате недовольства, обиды’; ***Наиболее предприим-***

чивые зрители *расставляют всех своих спутников и родственников во все очереди сразу*. [«Известия», 2001.06.18] — ‘с их стороны было предприимчиво расставить спутников во все очереди’.

При этом прилагательное может обозначать как актуальную характеристику на определенном интервале (*предприимчивые зрители*), так и постоянный признак (*болтливый губернатор*).

Итак, в режиме свободного денотативного присоединения прилагательное аналогично нерестриктивному относительному придаточному. И то и другое присоединяется к референтной определенной именной группе. Что касается обусловленного употребления прилагательного, то оно может быть аналогом и не-рестиктивного придаточного (в случае, например, причинного значения, ср.: *Фельдмаршал не посвятил в свои планы губернатора, который был слишком болтлив / поскольку тот был слишком болтлив*), и рестриктивного, если выражается значение условия, ср: *Испытывать весьма необычные ощущения при поездке в метро начнут скоро не слишком* ***дисциплинированные пассажиры***. [«Вечерняя Москва», 2002.04.11] — ‘если пассажиры недисциплинированны, начнут испытывать необычные ощущения’, ‘те пассажиры, которые недисциплинированны, начнут испытывать необычные ощущения’; *Судья вряд ли бы проронила ответную слезу и стала вникать в доводы* ***пристрастного свидетеля***. [Алексей Варламов. Купавна (2000)] ‘если свидетель пристрастен, его доводы не имеют ценности’; ‘судья не станет вникать в доводы такого свидетеля, который пристрастен’.

ПРИМЕЧАНИЕ: имя собственное.

В большинстве приведенных примеров существительное обозначает известный, уже введенный в рассмотрение объект, и в этом смысле аналогично имени собственному. Сами собственные имена, которые в диалогическом режиме, как известно, практически не присоединяют прилагательные, в режиме повествования подобного запрета не имеют. В случае «прилагательное + имя собственное» странно говорить, что прилагательное присоединяется к денотату «минуя семантику существительного», т. к. у имен собственных нет даже «обычной» лексической семантики, не говоря уже о встроенном предикате. Тем самым для прилагательных в контексте имен собственных возможно только денотативное присоединение.

Вот несколько примеров свободно присоединяемых прилагательных: ***Грозный Мишель Глотц*** *превратился в* ***тихого Мими***. [Сати Спивакова]; *Чтобы повару было сподручнее управиться с большой готовкой, прислали помощницу —* ***славненькую вертлявую Анечку***; *После истории у них была ещё химия.* ***Высокая надменная Антонина Михайловна****. Ногти как клыки у вампира*; *И руководит всем этим «городом» его чудо-мэр,* ***неутомимая Надежда Анатольевна Шагрова***; *В квартире стояла дубовая мебель и было даже пианино с хрустальными подсвечниками, измученное гаммами* ***усидчивой Людочки***. [И.Муравьева]; *Я всё хочу убедиться, что они живые, во плоти: молодой красавец Тьерри,* ***мужественная Алиса, добродушная Афра, резкая Хильда****…* [С. Юрский].

В группе имен собственных широко представлено также обусловленное употребление прилагательного: ***Маленький Валя Гафт****, отнюдь не отличавшийся на ниве успеваемости, имел потрясающую способность во время экзаменов концентрироваться и отвечать правильно* — ‘когда был маленьким’; *Артур бил уже несильно ладонями, точно пощёчины давал, наслаждаясь силой, из Гошиного носа брызнула кровь, а из глаз полились слёзы. На него было противно смотреть, и Колюня не жалел, что не вступился, а* ***умный Сережа****, мягко подталкивая, увёл подальше от дороги — не дай Бог кто из взрослых увидит* — ‘поскольку был умным; такое поведение есть проявление ума’; ***Аккуратный Вениамин*** *всегда имел в запасе несколько пар несношенных сапог, шинелей и множество белья*. [Э. Лимонов] — ‘поскольку был аккуратным, имел Р’; *Даже* ***терпимого Павла Алексеевича*** *он умел вывести из себя, и их встречи обыкновенно кончались ссорами, криками, хлопаньем дверями*. [Л. Улицкая] — ‘хотя был терпимым, выходил из себя’; *Обхаял* ***умнейших, талантливых Чубайса, Немцова, Гайдара, Кириенко***. [«Коммерсантъ-Власть», № 26, 1999] — ‘несмотря на ум и талант’.

3. Предикатное присоединение

В группе предикатного присоединения, как и в группе денотативного присоединения, есть разные модели связей прилагательного с контекстом.

3.1. Семантическое присоединение

В первой модели прилагательное непосредственно связано с внутренним предикатом существительного, который у большинства существительных рассматриваемой группы обозначает постоянную деятельность (*пастух*, *секретарь*) или единичное действие (*нарушитель*, *посетитель*). Назовем такое присоединение семантическим.

3.1.1. С функциональными существительными:

Многие ***авторитетные юристы*** *пытались прояснить это понятие, но все как-то путано получалось*. [«Арбитражный и гражданский процессы», 2004.10.25] — ‘имеют авторитет как юристы’; ***Квалифицированный юрист*** *смог бы само это заявление представить как деяние, образующее состав преступления*. [«Известия», 2002.10.13].

3.1.2. С реляционными существительными:

Что делать? Куда спрятаться от ***беспощадных врагов****? Но доктор не растерялся*. [К. И. Чуковский. Доктор Айболит (по Гью Лофтингу) (1929)]; *В эту минуту появился* ***таинственный незнакомец****. Ещё днём я заметил этого человека*. [Сергей Довлатов. Чемодан (1986)]; *Их сельский дом посещали лишь* ***самые близкие друзья*** *и дети от первых браков обоих*. [«Совершенно секретно», 2003.07.10], ср. также: *верный друг, надежный друг, искренний друг, любимый друг, преданный друг* и т. п.

3.1.3. С результативными существительными:

Так, в друзьях Советского Союза перебывали и «первый марксист Африки», а затем император-каннибал Бокасса, и угандийский тиран Иди Амин, и ***беспощадный убийца*** *из Эфиопии Менгисту Хайле Мариам*. [Александр Яковлев. Омут памяти. Т.1 (2001)]; *И вас, быть может, не удивит то странное на первый взгляд смешение* ***жестокого убийцы*** *с сентиментальным поэтом, пример которого мы видели на Лассенере*. [В. М. Дорошевич. Сахалин (Каторга) (1903)]; *Накапливание приводов могло стать причиной того, что* ***неисправимого нарушителя*** *в конце концов отправляли в исправительно-трудовые заведения*. [Алексей Козлов. Козел на саксе (1998)]; *Другое дело, что в войске не всегда находились* ***достойные исполнители*** *его прозорливых приказов, хотя и очень старались*. [Василь Быков. Главный кригсман (2002)] — либо ‘достойны того, чтобы их назначили исполнителями (достойны исполнять)’, либо ‘достойно (хорошо) исполняли’, но в обоих случаях прямая связь с предикатом существительного (‘исполнять’).

3.1.4. С актуальными существительными:

И все же суд состоялся. Пусть телевизионный, но с ***реальными свидетелями*** *реального преступления, настоящими адвокатами и прокурорами*. [«Известия», 2002.02.15] = ‘действительно видели’; ***Ценный свидетель*** *жив, это намного упрощало все дело*. [Татьяна Устинова. Большое зло и мелкие пакости (2003)] = ‘даст ценные показания как свидетель’.

Легко заметить, что почти все приведенные характеристики обозначают либо оценку, либо степень, т. е. выражают значение лексических функций BON / AntiBON или MAGN (в терминах модели «Смысл ⇔ Текст», см. [Апресян 1974], [Мельчук 1974]). Это, конечно, не случайно. Если определение относится не непосредственно к денотату (референту) существительного, а характеризует его как носителя функции, т. е. семантически связано с внутренним функциональным предикатом существительного, то это будут либо разнообразные относительные прилагательные, обозначающие разные аспекты самой деятельности (*нелегальный пассажир* — ‘нелегально едет’; *трехкратный победитель* — ‘три раза побеждал’; *потенциальный нарушитель* — ‘может нарушить’), либо качественные, которые так или иначе «усиливают» характеристику, заключенную в существительном (в противном случае они были бы денотативными).

Кроме того, семантическое присоединение не связано с режимом повествования, а, наоборот, предпочтительно в режиме сообщения.

3.2. Ситуативно связанное присоединение

Кроме таких «прямых» характеристик встречаются также «побочные» характеристики, когда прилагательное характеризует не то событие, которое заключено в названии лица, а другое, ситуативно связанное с ним или обусловленное им:

Как жаль, что царь после первого неудачного покушения и просьб помиловать ***неудачливых убийц*** *громко, на всю страну не сказал: «Даю шанс палачу промахнуться*!» [Фазиль Искандер. Понемногу о многом // «Новый Мир», 2000] — ‘не удалось убийство’; *Целые представления, занимательные, чуть не театральные зрелища, разыгрывались перед* ***восторженными зрителями***. [В. П. Карцев. Приключения великих уравнений (1970)] — ‘в восторге от увиденного’; *Некогда была опубликована расшифровка встречи Тарковского с внимательными,* ***заинтересованными зрителями****, собравшимися поговорить с создателем «Сталкера»*. [«Культура», 2002.04.08]; *Они раздражены, некоторые просто злы, недобро поглядывают на* ***радостного победителя*** — ‘радуется победе’.

Характер этой «ситуативной» связи может быть разным. В случае *неудачливый убийца* связь скорее «синтаксическая» (в том смысле, что она отражает отношение между неудачей и убийством, но, вообще говоря, вместо *убийцы* мог бы быть и *неудачливый ухажер*, и *неудачливый похититель*, и *неудачливый вратарь*), неудача может быть связана с разными видами деятельности и открывает широкий диапазон сочетаемости. Бывает семантически более тесная, импликативная связь: *Вот если бы в решении судьбы рукописи участвовал* ***прославленный автор*** *«Катюши», он, я думаю, отнесся бы к ней доброжелательно*. [«Вестник США», 2003.09.17] — не всякий автор становится знаменитым, но для *автора* (как, кстати, и для *вратаря*) это естественный и ожидаемый признак (‘если X сочинил Y, он может прославиться’), — в отличие от *ухажера* или *убийцы*, для которых слава хотя и не исключена, но не входит в набор «нормальных ожиданий».

Употребление предикатно присоединяемых прилагательных, так же как и денотативно присоединяемых, может быть обусловленным:

И в том и в другом спектакле было много песенок, «танчиков», дешевых хохм, на которые охотно клевал ***невзыскательный зритель****, заполнивший*

партер и ярусы театра Маяковского. [Михаил Козаков. Актерская книга (1978–1995)] — ‘клевал, поскольку был невзыскателен как зритель’; *Но это, возможно, какое-то время останется для* ***неискушенного зрителя*** *незамеченным*. [«Ландшафтный дизайн», 2002.03.15] — ‘если зритель неискушенный, не сразу заметит’; — *Сейчас можно найти такого* ***умного зрителя****, о котором вы говорили*? [«Богатей» (Саратов), 2003.05.22] — ‘ведется поиск не любого зрителя, а умного’; *Нужны компетентные,* ***честные исполнители***. [А. И. Деникин] — ‘нужны честные, а не любые’.

4. Заключительные замечания

Денотативное присоединение (аналог нерестриктивного относительного придаточного) — это одно из проявлений механизма компрессии, когда к предложению присоединяется дополнительная предикация в качестве второстепенной. Принципы денотативного присоединения в случае одного прилагательного или целого придаточного похожи: предикация, выражаемая атрибутом, должна быть актуально связана с интервалом основной ситуации: *Пришел сосед, который рассказал последние новости* (пришел и рассказал — непосредственное следование во времени). При денотативном присоединении прилагательного этот принцип также является основным: прилагательное должно обозначать признак, актуальный на интервале основной ситуации. Это, в свою очередь, определяет (пусть и не жесткие) требования к прилагательному. Наиболее характерными прилагательными при денотативном присоединении являются прилагательные со значением поведения — либо актуального (*Вежливая продавщица подала еще один экземпляр*), либо постоянного поведенческого признака (*С ними приезжал вежливый адвокат*).

Ясно, что не все существительные одинаково уместны в нарративном режиме. Так, номинация *друг* с денотативно присоединенным прилагательным обычно не обозначает друга говорящего и вообще ведет себя весьма прихотливо (иногда, например, — по-разному в единственном и множественном числе).

Как показывает материал, денотативное присоединение есть у всех типов существительных, что вполне логично, т. к. если прилагательное семантически не связано с существительным, то и не важно, к какому классу относится существительное. Поскольку между существительным и прилагательным нет семантической связи, нельзя сформулировать закономерностей выбора прилагательного и сочетаемости с существительным. Это не означает, однако, что имена, формально образующие атрибутивную конструкцию, выбираются произвольно. Выбор существительного ситуативно обусловлен: если речь идет о ситуации купли-продажи, уместно обозначение *продавщица*, если о путешествии — *пассажир*. Эти обозначения задают интервал, а прилагательные сообщают признаки денотата, релевантные на этом интервале (так обстоит дело в режиме повествования, т. е. «изображения», в режиме сообщения это признаки, важные для автора сообщения).

В связи с таким поведением прилагательных возникает множество вопросов, которые мы здесь сможем только обозначить.

Общей презумпцией (допущением) при создании синтаксических конструкций (словосочетаний, предложений) из слов является та, что синтаксические связи возникают на основе семантических. Даже если формального выражения синтаксической зависимости нет (как в случае так называемого примыкания, ср. ***Медленно*** *он вообще никогда не* ***ходит*** — в данном примере нет даже контактного расположения), все равно обнаруживается семантическая связь (*ходить* → ‘движение’ → ‘скорость’ → *медленно*). Этот принцип должен, естественно, распространяться и на атрибутивные сочетания: *горячий чай* (у чая есть параметр «температура»), *узкая тропинка* (у тропинки есть параметр «ширина»).

Однако, если последовательно рассмотреть разные случаи атрибутивных конструкций, то окажется, что очень часто семантическое согласование (т. е. семантическое присоединение) отсутствует. Во-первых, сразу же выпадают относительные прилагательные: они присоединяются на основе совершенно других связей — актантных (*рыбный торговец* — здесь никакого семантического согласования между человеком и рыбой не требуется, т. к. на семантическом уровне обнаруживается другая связь — управление: ‘торгует рыбой’), обстоятельственных (*ночной пассажир* — ‘ехал ночью’). Кроме того, аналогичные модели присоединения встречаются и среди качественных прилагательных. Например, оценка приписывается не «прямо», а «извне», через посредника: *невкусный чай* — субъект оценки попробовал чай и приписал ему оценку «от себя», не характеризуя прямо его денотативных свойств. Возникает вопрос, есть ли у не-личных (предметных) существительных (*нож*, *рыба*, *чай* и т. п.) механизм денотативного присоединения не-оценочных качественных прилагательных и можно ли (стоит ли) в этом случае говорить о семантической сочетаемости.

В целом, подобный материал заставляет заново рассмотреть (и, возможно, заново сформулировать) принципы сочетания слов в предложении.

Вообще, поведение прилагательных в составе номинаций — мало исследованная тема. Отдельные ее аспекты рассматривались в теории референции (см., напр., [Гак 1998], [Вежбицкая 1982]), но именно с точки зрения референциальных статусов и функ-

ций. Между тем есть множество других, не менее важных аспектов, касающихся принципов встраивания прилагательного в семантическую структуру предложения (вклад прилагательного в сообщаемую информацию, установление его статуса как денотативно или семантически присоединяемого, установление семантических связей, нередко не совпадающих с поверхностно-синтаксическими). Исследования в этом направлении, во-первых, обогатят наши представления о принципах построения и интерпретации текстов, «упаковки» информации, стратегий воздействия на слушателя / читателя, а с другой стороны, позволят уточнить семантические классификации существительных и прилагательных, что весьма актуально в связи с широким распространением семантической разметки электронных словарей и корпусов, разработкой и тренировкой программ автоматического анализа текста.

Литература

1. *Апресян Ю. Д.* Лексическая семантика. Синонимические средства языка. М.: 1974.
2. *Арутюнова Н. Д.* Предложение и его смысл. М.: 1976.
3. *Вежбицкая А.* Дескрипция или цитация // Новое в зарубежной лингвистике. Вып. 13. М.: 1982.
4. *Гак В. Г.* Типология лингвистических номинаций // В. Г. Гак. Языковые преобразования. М.: 1998.
5. *Мельчук И. А.* Опыт теории лингвистических моделей «Смысл ⇔ Текст». М.: 1974.
6. *Падучева Е. В.* Семантика нарратива // Е. В. Падучева. Семантические исследования. М.: 1996.
7. *Шмелев А. Д.* Русский язык и внеязыковая действительность. М.: 2002.
8. *Шмид В.* Нарратология. М.: 2004.

Архитектура Системы охвата информационных связей объектов мониторинга

Architecture of system of coverage of informative connections of monitoring objects

Ландэ Д. В. (dwl@visti.net), **Брайчевский С. М.** (smb@visti.net), **Дармохвал А. Т.** (hval@visti.net), **Жигало В. В.** (vladlen@visti.net)

Информационный центр «ЭЛВИСТИ», Киев, Украина

Представлен подход к построению полнотекстовой информационно-поисковой системы, основными элементами которой являются не отдельные термы, а взаимосвязи между понятиями, экстрагируемыми из текстовых документов. Описаны основные компоненты и архитектурные решения, применяемые в данной системе, а также интерфейс пользователя.

В настоящее время информационное пространство представляет собой динамическую среду, наполнение которой постоянно изменяется. Существующие доступные фактографические базы данных структурированной информации не всегда могут прийти на помощь исследователю-аналитику. Для оперативного определения фактов и сущностей, моделирования информационных связей между ними наиболее перспективным подходом оказывается учет информации, знаний, которые содержатся в неструктурированных текстовых документах, в частности, в Интернет. Поиск в базах данных неструктурированной текстовой информации может применяться для задач наведения исследователей-аналитиков «на цель» в условиях, когда фактографические базы данных структурированной информации труднодоступны, неполны, неоперативны.

Неструктурированные тексты содержат в себе несравненно больше важной информации, чем структурированные записи баз данных, именно в силу того, что формализации подлежит сравнительно небольшой сегмент информации. В настоящее время появляется все больше качественных инструментальных средств извлечения понятий из неструктурированных текстов [1], [2], [3].

Сегодня ни у кого не вызывает сомнений то, что не удастся создать единую универсальную информационно-поисковую систему, которая решала бы одинаково эффективно все поисковые задачи. Поэтому представляется актуальным поиск технологических решений, ориентированных на применение к различным предметным областям, предполагающим определенные информационные потребности.

Ниже будет представлено одно из решений, основанное на использовании в процессе полнотекстового информационного поиска взаимосвязей между понятиями.

В настоящее время, когда у пользователей уже накоплен большой опыт работы с традиционными информационно-поисковыми системами, оказалось очевидным, что факты или понятия, которые ищутся с помощью таких систем, сами по себе зачастую бессмысленны. Например, если пользователя интересуют информационные связи Сбербанка России с другими банками или частными лицами, то он не знает, какие банки или фамилии ему указать в запросе, а все документы, содержащие словосочетание «Сбербанк России» физически невозможно. В таких случаях информационные связи, интенсивность которых выходит за рамки статистического фона, как правило, отражают реальность.

Интерпретируют обычно не сами понятия или факты, а взаимосвязи между ними. «..Важным оказывается не столько исследование самих понятий, сколько исследование их взаимосвязи. Именно взаимосвязь способствует пониманию мотивационно-целевых особенностей отношений человека...» [4] . То есть пользователя интересует не понятие само по себе, а понятие в окружении, чтобы сразу иметь представление о предметной области, при необходимости направить уточняющий поиск в нужном направлении. Элементы такого подхода можно видеть, например, в «облаках» системы Quintura (http://quintura.ru), но там отображаются не понятия/сущности, а наиболее часто используемые термы.

Таким образом объективно существует необходимость построения эффективной полнотекстовой

информационно-поисковой системы, обеспечивающей поиск не по отдельным термам или понятиям, а по взаимосвязям между сущностями, присутствующими в документах, то есть создания систем, которые будем условно называть «базами данных связей» (БДС).

База данных практически любой традиционной информационно-поисковой системы может быть рассматриваться в виде графа, вершинами которого выступают объекты — термы, понятия, дескрипторы и др., а ребрами — их связи. Вместе с тем, основа поиска в этих случаях — поиск вершин, то есть поиск объектов. Поиск по взаимосвязям, ребрам, кажется на первый взгляд менее эффективным. Действительно, если предположить, что в графе N вершин, то ребер теоретически может составлять $N(N\text{-}1)/2$, то есть, если предположить, что вершин всего 100 тыс., то ребер может оказаться около 5 млрд, что соответствует достаточно большой базе данных даже по современным понятиям. Вместе с тем, если в качестве вершин графа использовать такие понятия, как имена людей и названия компаний из новостных документов, то оказывается, что соответствующая матрица инцидентности оказывается очень разреженной. Измерения показали, что при количестве отдельных понятий, извлеченных из 5 млн. новостных документов, равном примерно N = 1,5 млн., количество связей составило всего лишь v = 4 млн., то есть коэффициент разреженности матрицы данного графа составил:

$$K = \sqrt{\frac{v}{N(N-1)/2}} \approx 0,002.$$

Кроме того, как показали эксперименты, распределение степени вершин в подобных графах — степенное (см. рис. 1) [5], что свидетельствует о так называемой безмасштабности, то есть о том, что многие характеристики (в частности, соотношение количества вершин и ребер), должно оставаться на одном уровне. Поэтому в качестве основы построения базы данных связей сегодня оказывается технически возможным использование ребер рассматриваемого графа — связей между отдельными понятиями.

Исходя из результатов исследований, была создана база данных связей объектов путем мониторинга интернет-ресурсов. Пользователи этой системы фактически получают доступ к базе данных информационных взаимосвязей интересующих их объектов. Под информационной взаимосвязью в узких рамках представленной ниже системы понимается совместное упоминание объектов в некоторой информационной единице, принятой в качестве основной. Такими единицами могут быть документы, разделы документов, абзацы и т. д. Описываемая ниже реализация охватывает обычные статистические связи, однако, следует отметить, что в рамках предлагаемого подхода взаимосвязи могут быть определены и другими способами; в этой работе не идет речь о создании средств выделения понятий из текстов и установления связей между ними. Предлагается подход к созданию ИПС, в которой подобные средства играют важную, но вспомогательную роль. То есть в рамках данной работы речь не идет о создании высокоэффективной системы выявления связей, которая обладала бы явными преимуществами по сравнению с другими подобными системами. Вместе с тем речь идет о поисковой системе, ориентированной на использование конечным пользователем в режиме онлайн.

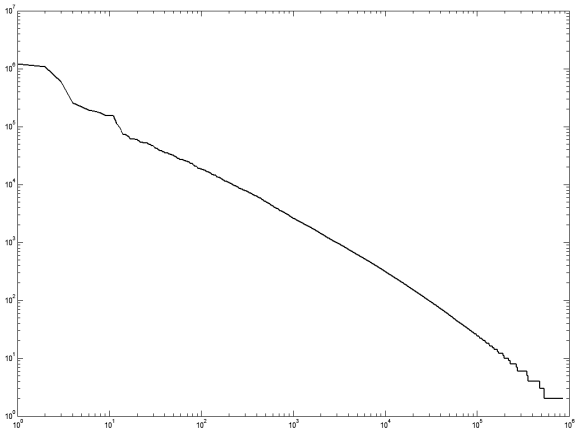


Рис. 1. Ранжированный график степеней вершин графа понятий (количества исходящих ребер) в логарифмической шкале

В качестве массивов новостной информации использовались фрагменты базы данных системы контент-мониторинга InfoStream [6], а также результаты мониторинга специализированных вебслужб, таких как базы данных биографий людей, организаций, служб трудоустройства и т. п.

Информационные взаимосвязи между понятиями выявляются путем обработки текстовых массивов и хранятся в специальной базе данных. Набор понятий, используемый при построении БДС, формируется путем экстрагирования данных из того же текстового массива, что придает системе целостность.

В корпоративной информационной инфраструктуре база данных связей может использоваться различным образом, например, отдельно, либо возможности БДС могут быть дополнены возможностями существующих полнотекстовых и/или фактографических баз данных (рис. 2). При этом основным результатом работы БДС является построение карт связей, а в качестве побочного эффекта, реализующего «режим доказательства», может рассматриваться извлечение самих документов как источников связей.

При проектировании БДС использовались решения, которые можно отнести к самым пер-

спективным в области создания информационно-аналитических систем, в частности, теория и технологии глубинного анализа тестов — Text Mining [7], в том числе развитая методология экстрагирования понятий [1, 2, 8], теория и технологии баз данных сверхбольших объемов, концепция «сложных сетей» (complex networks) [9, 10]. Теория сложных сетей изучает характеристики, учитывая не только на топологию сетей, но и статистические феномены, распределение весов отдельных вершин (в качестве которых можно рассматривать сущности, понятия, факты) и ребер, эффекты протекания и проводимости в сетях и т. п.

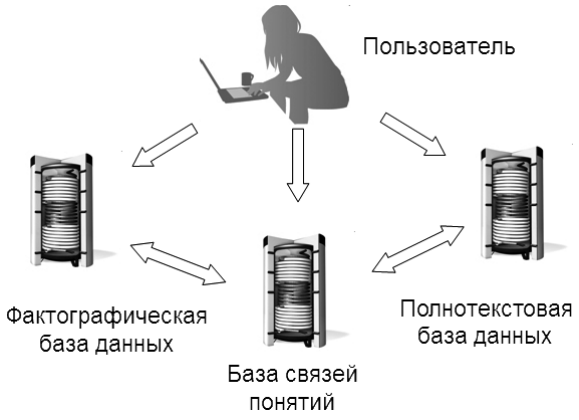


Рис. 2. Место базы данных связей понятий в корпоративной информационной инфраструктуре

На рис. 3 схематически представлены основные технологические этапы формирования базы данных связей. С помощью программы-робота осуществляется сканирование выбранных веб-ресурсов, которые содержат информацию, относящуюся к объектам исследований. После этого осуществляется экстрагирование необходимых пользователям понятий, например, имен персон, наименований брендов, компаний, электронных адресов и т. п. Отобранные понятия и соответствующие отношения между ними, загружаются в базу данных связей, которая также содержит ссылки на документы-первоисточники. Средства экстрагирования понятий ориентированы на обработку документов, сканируемых из Интернет, представленных как на русском, украинском, так и на английском языках. Реализовано автоматическое функционирование системы в режиме мониторинга интернет-ресурсов по мере их поступления.

Для настройки средств экстрагирования понятий предусмотрены таблицы шаблонов, такие как таблица фамилий известных персон, слов, заведомо не являющихся фамилиями (стоп-словарь), возможных имен, изменяемых окончаний и соответствующих им нормальных форм. Для экстрагирования названий компаний предусмотрены такие таблицы, как названия известных компаний, «префиксов», используемых для выявления неизвестных заранее компаний (применяется для русско- и украиноязычных документов), например, «АО», «ООО», «ТОВ», «АОЗТ» и др., «суффиксов», используемых для выявления неизвестных заранее компаний (применяется для англоязычных документов), например, «Ltd», «Inc», «Corp» и др.

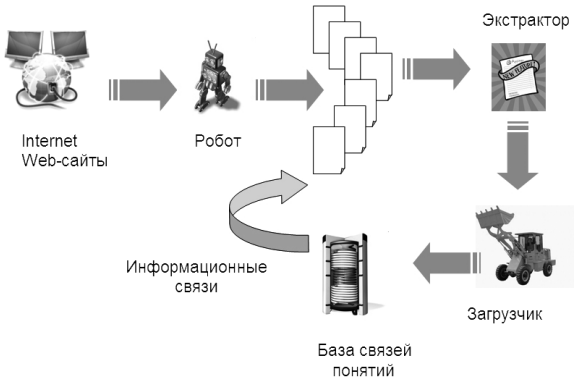


Рис. 3. Схема формирования базы данных связей

Предложенный подход к поиску, естественно, влечет за собой некоторые особенности в реализации архитектуры базы данных связей понятий. Кроме того, архитектура БДС должна быть ориентирована на такие возможные применения, как выявление неявных связей (не выявленных явно комплексом экстрагирования понятий), поиск отдельных объектов, а также взаимосвязь с существующими фактографическими базами данных. Вариант такой архитектуры в настоящее время реализован и используется в качестве компоненты системы конкурентной разведки X-Files.

Сама база данных связей состоит из двух разделов:

1. CTU (Concept-Time-Unq. number), отражающего появление понятия в текстах документов, основная таблица которого содержит такие поля, как понятие, дата, время упоминания, уникальный номер документа, абзаца, предложения.
2. CCN (Concept-Concept-Number), отражающего совместное появление пары понятий в текстах.

Информационные взаимосвязи объектов мониторинга расширяется за счет взаимодействия с фактографическими базами данных. Расширенная система фактически включает две сети. Первая сеть условно называется «Картой явных связей», а вторая, соответственно, «Картой вероятных связей» (рис. 4).

Для построения карты явных связей используется наложение двух промежуточных сетей (или слоев) — информационной и фактографической. Информационная сеть формируется из понятий,

экстрагируемых из текстовых документов, а фактографическая — на основании фактографических данных. Важная проблема связана с необходимостью выявления неочевидных закономерностей и связей понятий. На сегодня известно несколько путей решения этой проблемы, например, на основе концепции сложных сетей [11]. В настоящее время авторами проводятся исследовательские роботы по формированию карты вероятных связей, которые базируются на подходе, описанном в [12].

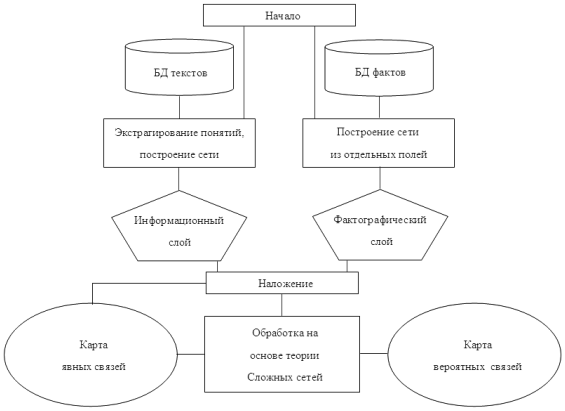


Рис. 4. Схема формирования карт явных и вероятных связей

Система позволяет пользователю в онлайн-режиме получать карты связей (КС) для указанных им объектов и помогает интерпретировать результаты. Предусматривается, что пользователь вводит в качестве запроса системе объект. Запрос направляется к БДС, откуда выбираются соответствующие ему фрагменты — карты связей (уровень детализации и временная ретроспектива должны указываться параметрически).

После выявления релевантных объектов и связей выполняются процедуры их автоматической группировки (кластеризации) и визуализации, результаты предъявляются пользователю в виде КС. Карты связей представляются в нескольких форматах, в том числе в табличном (таблица взаимосвязей понятий), графическом (круговая диаграмма), в виде динамических Java-диаграмм (графов связей), построенных с помощью средств TouchGraph.

Интерфейс взаимосвязей субъектов позволяет пользователю выбрать:

- вид субъекта;
- наименование субъекта;
- промежуток времени;
- глубину детализации.

Граф связей (ГС) строится с помощью апплетов Java и представляет собой графический объект, который содержит в своем составе узлы и ребра. Каждый элемент ГС имеет контекстное меню, которое является дополнительным элементом управления в интерфейсе пользователя БДС (рис. 5).

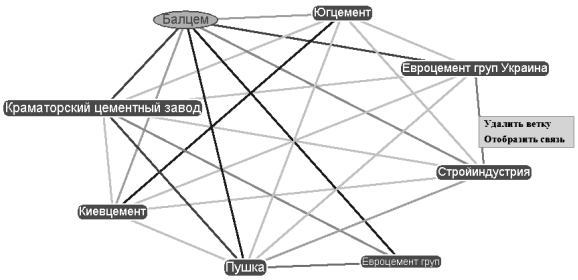


Рис. 5. Контекстное меню ребра

Объекты, которые имеют больше количество связей, изображаются с помощью большего шрифта. Ребра, соответствующие большему количеству связей, изображаются более темными линиями. Построенная сеть имеет собственные средства управления: изменение масштаба (с помощью меню «масштаб» или полосы прокрутки в верхней части экрана); перемещение всего графа; перемещение объекта; изменение конфигурации; подсветка связей выбранного узла и т. п.

В *таблице взаимосвязей понятий* данные представлены в виде квадратной таблицы, строки и столбики которой соответствуют объектам, связанным с исходным (рис. 6). Ячейки таблицы окрашивающиеся разными оттенками серого цвета, в зависимости от количества взаимосвязей объектов, которым соответствует строка и столбец. При наведении на ячейку таблицы указателя «мыши» на экран выводится количество взаимосвязей данной пары объектов.

Фирмы		1	2	3	4	5	6	7	8	9	10	11	12
Балцем	1												
Евроцемент груп-Украина	2												
Краматорский цементный завод Пушка	3												
Стройиндустрия	4												
Бахчисарайский комбинат Стройиндустрия	5												
Укрцемент	6												
Киевцемент	7												
Юцемент	8												
Dyckerhoff AG	9												
Волынь-Цемент	10												
Николаевцемент	11												
Альтцем	12												

Рис. 6. Таблица взаимосвязей понятий

Круговая диаграмма представляет собой графический объект, в котором узлы (все кроме главного равномерно распределены на кругу, а основной занимает центральное положение) соответствуют отобранным объектам, а ребра — связям между ними. На круговой диаграмме (рис. 7) ребра, соответствующие большему количеству связей, изображаются более толстыми и темными линиями. При нажатии на узел круговой диаграммы кнопкой «мыши» открывается список документов, в которых упоминается выбранный объект совместно с основным.

Приведем еще один пример использования БДС, случай, когда пользователя интересуют информационные связи Сбербанка России. Разумеет-

ся, для запроса «Сбербанк России» может быть выявлено множество различных связей, но при этом существует простой и надежный критерий ранжирования результатов, состоящий в отсечении статистического фона. В рассматриваемом случае, задав соответствующий запрос можно получить граф (рис. 8) наиболее связанных со Сбербанком России объектов (персон и компаний). И если нахождение фамилий руководителей банка (председателя правления, первого заместителя председателя правления и руководителя дочернего банка) является достаточно очевидным результатом, то связи между отдельными банками позволили выявить (после обращения к документам-первоисточникам) неочевидные на первый взгляд факты, например, то, что УкрСиббанк и УкрСоцбанк являются банками-партнерами.

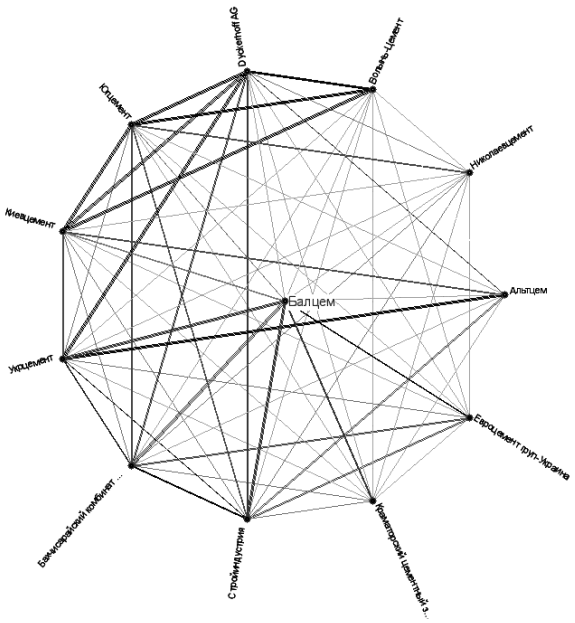


Рис. 7. Круговая диаграмма

Представленный подход может рассматриваться как основа построения информационно-поисковых систем, в которых изначально решены вопросы оперативности, отсеивания информационного шума. Рассматриваемая реализация — БДС имеет свойство масштабирования по трем параметрам: объему баз данных, составу понятий, которые используются, и по инфраструктурному окружению.

К настоящему времени еще не получено количественных оценочных характеристик для реализованного варианта системы, возможно при распространении представленного подхода, поиск по связям станет одной из дорожек TREC или РОМИП, как это происходило с подключаемыми к ней компонентами. Отдельные компоненты описанной системы, такие как модули экстрагирования и определения взаимосвязи понятий, средства визуализации и кластеризации, выявления «неявных связей», и т. п. могут заменяться и привязываться к информационно-поисковой системе с помощью прикладных программных интерфейсов.

Анализируя связи в сети, можно определить многие неочевидные свойства, например, выявить наличие кластеров, определить их состав, различия в связности внутри и между кластерами, идентифицировать ключевые элементы, которые связывают кластеры между собой и т. п. Серьезным препятствием при анализе является неполнота информации о связях между отдельными узлами сети. Вместе с тем сегодня уже существуют алгоритмы [11, 12], с помощью которых становится возможным с высокой вероятностью восстановить отсутствующие фрагменты связей. Даже не имея полного описания информационной сети, можно получать репрезентативную выборку «реальных» связей и по ней достроить всю сеть. Перспективы развития созданной системы — усложнение учитываемых связей, учет семантики контекста понятий в документах при их экстрагировании, отбор перечня действительно полезных баз данных текстовых документов, учет большего количества сущностей (понятий).

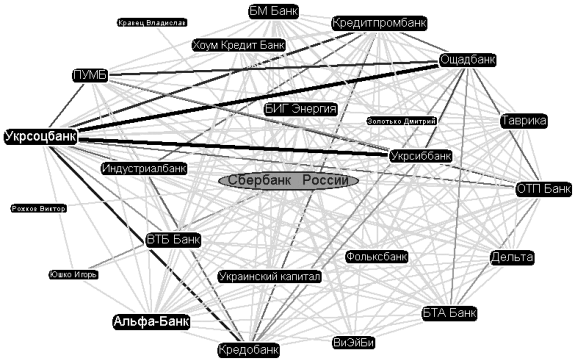


Рис. 8. Граф информационных связей понятия «Сбербанк России»

Представленная система является связующим звеном связи между полнотекстовыми и фактографическими базами данных. Очевидно, что реальный прорыв в области информационно-аналитической работы возможный лишь в результате агрегирования разных направлений. Базирующиеся на нескольких конкурирующих ранее точках зрения подходы на сегодня могут рассматриваться как пути создания современной мощной информационно-аналитической системы.

Литература

1. *Grishman R.* Information extraction: Techniques and challenges. In Information Extraction (International Summer School SCIE-97). Springer-Verlag, 1997.
2. *Гершензон Л. М., Ножов И. М., Панкратов Д. В.* Система извлечения и поиска структурированной информации из больших текстовых массивов СМИ. Архитектурные и лингвистические особенности // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог'2005» (Звенигород, 1–6 июня, 2005 г.)/ Под ред. И. М. Кобозевой, А. С. Нариньяни, В. П. Селегея. — М.: Наука.
3. *Додонов А. Г., Ландэ Д. В.* Выявление понятий и их взаимосвязей в рамках технологии контент-мониторинга // Регистрация, хранение и обработка данных, 2006, Т. 8, № 4. — С. 45–52.
4. *Массон Г. В.* Взаимосвязь системы личностных терминальных ценностей и типов межличностных отношений: Дис. ... канд. психол. наук : 19.00.01: Красноярск, 2004. — 146 с. РГБ ОД, 61:05–19/11
5. *Ландэ Д. В., Григорьев А. Н., Брайчевский С. М., Дармохвал А. Т., Снарский А. А.* Труды 9-ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2007, Переславль-Залесский, Россия, 2007. — С. 148–150.
6. *Григорьев А. Н., Ландэ Д. В.* Адаптивный интерфейс уточнения запросов к системе контент-мониторинга InfoStream // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог'2005» (Звенигород, 1–6 июня, 2005 г.)/ Под ред. И. М. Кобозевой, А. С. Нариньяни, В. П. Селегея. — М.:Наука — С. 109–111.
7. *Berry M. W.* Survey of Text Mining. Clustering, Classification, and Retrieval. — Springer-Verlag, 2004. — 244 p.
8. *Ермаков А. Е.* Автоматическое извлечение фактов из текстов досье: опыт установления // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2007» (Бекасово, 30 мая — 3 июня 2007 г.) / Под ред. Л. Л. Иомдина, Н. И. Лауфер, А. С. Нариньяни, В. П. Селегея. — М.: Изд-во РГГУ, 2007. — С. 172–177.
9. *Newman M. E. J.* The structure and function of complex networks // SIAM Review. — 2003. — Vol. 45. — pp. 167–256.
10. *Ландэ Д. В., Снарский А. А., Безсуднов И. В.* Интернетика: Навигация в сложных сетях: модели и алгоритмы. — М.: Либроком (Editorial URSS), 2009. — 264 с.
11. *Clauset A., Moore C., Newman M. E. G.* Hierarchical structure and the prediction of missing links in networks // Nature. — 2000. — Vol 453. — pp. 98–101.
12. *Снарский А. А., Ландэ Д. В., Женировский М. И.* Метод выявления неявных связей объектов // Труды 11й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL'2009, Петрозаводск, Россия, 2009. — С 46–49.

Влияние паузы хезитации на понимание синтаксической структуры предложения носителями русского языка[1]

Effects of hesitation in speech on syntactic structure in comprehension: evidence from russian speakers

Лауринавичюте А. К. (annlaurin@gmail.com),
Федорова О. В. (olga.fedorova@msu.ru)

Московский государственный университет имени М. В. Ломоносова

Исследовалось влияние паузы хезитации на понимание сложноподчиненных предложений с союзными словами *перед* и *после* носителями русского языка. Выяснилось, что при наличии паузы перед союзным словом сложные конструкции представляют меньше трудностей, а более простые понимаются хуже; наличие паузы после союзного слова не влияет на понимание.

Введение

Наша спонтанная речь редко бывает идеальной — мы то и дело или запинаемся, или тянем и даже повторяем отдельные слова, или исправляем неудачно начатую фразу. Долгое время подобного рода **речевые сбои** было принято игнорировать, однако в последние десятилетия интерес к ним неуклонно растет. Начиная с 60-х годов прошлого века (обзор самых ранних работ см. в [Николаева 1970]) данные исследования проводятся в области (социо) лингвистического анализа бытового диалога (напр., [Fox et al. 1996]), (психо)лингвистического изучения спонтанной речи (см., в частности, работы Г. Кларка ([Clark & Wasow 1998], [Clark & Fox Tree 2002]) и материалы четырех прошедших с 1999 года семинаров по речевым сбоям в спонтанной речи; следующий подобный семинар пройдет в 2010 году в Японии (http://cogsci.l.chiba-u.ac.jp/diss-lpss2010/)), компьютерного моделирования [Ferreira & Bailey 2004], а также (психо)лингвистического экспериментального подхода ([MacGregor et al. 2009]).

Предметом изучения в настоящей работе будет частный вид речевых сбоев, а именно, так называемые **паузы хезитации** (паузы колебания), то есть некоторый перерыв в фонации (для того, чтобы мы могли воспринять его на слух, его длительность должна быть больше 150–200 мс), часто заполненный различными звуками. Заполненные или незаполненные паузы хезитации обычно свидетельствуют о том, что следующий за текущим фрагмент высказывания по каким-то причинам еще не готов к артикуляции и говорящему требуется дополнительное время на его формулирование.

Настоящая работа выполнена в рамках **психолингвистического подхода**, поэтому далее в разделе 1 мы более подробно остановимся на особенностях этого подхода применительно к исследованию хезитационных пауз. В разделе 2 мы покажем, каким образом наличие паузы хезитации перед началом артикуляции высказывания влияет на его понимание.

1. Паузы хезитации с точки зрения психолингвиста

В отличие от исследователей, работающих в области изучения спонтанной речи, психолингвисты рассматривают паузы хезитации преимущественно с точки зрения слушающего, то есть изучают, как паузы влияют на языковое поведение испытуемых в процессе **понимания речи**. Существует несколько разных методик, которые при этом традиционно используются, среди них есть как онлайновые (ко-

[1] Работа выполнена при частичной финансовой поддержке РГНФ (проект № 08-04-00165а). Авторы выражают благодарность А. А. Кибрику, Е. В. Печенковой и В. И. Подлесской за критические замечания и советы, высказанные при подготовке работы.

торые позволяют исследовать действие языковых механизмов в режиме реального времени), как то: регистрация движений глаз, вызванные потенциалы мозга, так и оффлайновые (используя которые исследователь изучает языковое поведение испытуемых уже после завершения самого процесса понимания): разыгрывание сцен, раскладывание картинок. Более подробное описание оффлайновых методик будет дано в разделе 2, а сейчас мы приведем несколько примеров онлайновых исследований.

В эксперименте с использованием методики **регистрации движений глаз** [Arnold et al. 2003] было продемонстрировано, что наличие паузы хезитации в инструкции способствует тому, что испытуемые фиксируют свой взгляд на новом объекте (а не на том, который был уже упомянут ранее) еще до того, как произнесено само слово, называющее этот объект. В случае, если паузы хезитации не было, испытуемые смотрели на тот же объект, с которым они работали ранее. В работе [Corley et al. 2007] был использован метод **вызванных потенциалов мозга**. Известно, что когда человек слышит слово, которое он не ожидает услышать в этом контексте, у него наблюдается так называемый эффект N400 (негативная реакция через примерно 400 мс после начала звукового сигнала). Авторы работы обнаружили, что данный эффект становится значимо меньше, если перед этим неожидаемым словом испытуемые слышат хезитационную паузу. Таким образом, паузы хезитации действительно оказывают влияние на синтаксическую обработку высказывания в процессе понимания. Как нам представляется, функция паузы хезитации состоит в том, чтобы сигнализировать отключение дальнейшего развития высказывания «по умолчанию». То есть, если слушающий настроен на то, что следующее слово будет определенным (и он знает или подозревает, каким именно), наличие паузы сообщает ему, что эта модель не будет работать, и ему, возможно, нужно обратить больше внимания на последующий отрывок дискурса, чтобы суметь построить правильную модель ситуации (situational model).

Важным теоретическим вопросом в данной области является вопрос о том, использует ли говорящий паузы хезитации намеренно, чтобы сигнализировать слушающему о возникших у него языковых сложностях (в частности, авторы известной работы [Clark & Fox Tree 2002] считают, что говорящий может, хотя и не полностью, контролировать возникновение хезитационных пауз), или же эти паузы возникают автоматически, независимо от воли говорящего, являясь как бы дополнительным эффектом возникающих трудностей при порождении. Как полагают авторы недавней работы [Corley & Stewart 2008], на этот вопрос еще нет однозначного ответа. Отметим тем не менее, что не подлежит сомнению то, что иногда говорящий намеренно использует паузы хезитации как особый риторический прием [Шмелев 2005].

В любом случае, однако, общепризнанным сейчас считается тот факт, что паузы хезитации возникают в случае наличия у говорящего определенных сложностей в планировании текущего высказывания — чем больше этих сложностей, тем более вероятно появление паузы хезитации и тем больше ее продолжительность. В дальнейшей работе нас будут интересовать собственно языковая сложность высказывания, которое говорящий собирается произнести, а не другие (в том числе и внеязыковые) проблемы, влияющие на процесс коммуникации. Для экспериментальной проверки положения о том, что языковая сложность влияет на возникновение у говорящего паузы хезитации, которая, в свою очередь, влияет на интерпретацию этого высказывания слушающим, нам нужно более строго сформулировать понятие языковой сложности, что само по себе представляет нетривиальную задачу. В нашем эксперименте для этих целей мы будем исследовать конструкции с различной степенью **синтаксической сложности**, что уже не раз было использовано в психолингвистической практике.

Так, в работе [Clark & Wasow 1998] авторы, опираясь на анализ спонтанной речи, показали, что чем сложнее синтаксическая структура высказывания, тем более вероятно наличие паузы хезитации перед началом его артикуляции. Данный вывод был подтвержден экспериментально в работе [Ferreira 1991], в которой автор варьировал сложность синтаксической структуры предложений, которые испытуемым нужно было произнести. Оказалось, что предложение типа *The river near their city empties into the bay that borders the little town* требовало большего времени на подготовку, чем предложение *The large and raging river empties into the bay that borders the little town,* но меньше, чем предложение *The river that stopped flooding empties into the bay that borders the little town*.

2. Исследование пауз хезитации на материале русского языка

В работе [Подлесская & Кибрик 2009], основанной на корпусе спонтанной русской речи, было показано, что именно **заполненные паузы** являются типичным маркером хезитации. В среднем подобные речевые сбои встречаются в этом корпусе 1,9 раза на каждые 100 слогов, что, как отмечают авторы, хорошо согласуется с обобщенными данными из работы [O'Connell & Kowal 2004]. В нашем эксперименте мы решили использовать заполненные паузы типа мэканья средней продолжительностью 550 мс, перед которыми имеется также небольшая (до 300 мс) незаполненная пауза.

2.1. Стимульный материал эксперимента

В качестве стимульного материала были выбраны **сложноподчиненные предложения с придаточными времени**, которые отличаются друг от друга сразу по нескольким основаниям. С точки зрения синтаксиса, возможно два порядка следования клауз: главное предложение предшествует придаточному, примеры (3) и (4), или следует за ним, (1) и (2). С точки зрения семантики союзного слова, предложение с *перед тем как* описывает ситуацию, в которой действие главного предложения предшествует действию придаточного, в то время как предложение с *после того как* описывает обратную ситуацию. С точки зрения прагматики, в (2) и (3) порядок упоминания событий соответствует порядку их реального проистекания, а в (1) и (4) — нет.

Изучение данного феномена на английском языке началось с работы [Clark & Clark 1968], в которой было показано, что взрослые испытуемые делают больше ошибок в случае, если (i) порядок упоминания событий не соответствует их естественному порядку (**гипотеза 1**). Вскоре после этого появились две другие гипотезы: лучше понимаются те предложения, в которых (ii) содержится союз *before*, а не *after* (**гипотеза 2**), и (iii) сначала идет главное предложение, а потом зависимое (**гипотеза 3**).

Исследование на материале русского языка было описано в работе [Федорова 2005]. Эксперимент был проведен по методике разыгрывания сцен: испытуемые должны были перекладывать, пользуясь компьютерной мышью, объекты на экране в соответствии с инструкциями, которые они слышали в наушники. В результате оказалось, что предложения типа (3) *А теперь переставьте коричневый диван под змею, перед тем как положите голубую птицу в корзинку* вызывают у испытуемых намного больше ошибок, чем во всех остальных случаях, вместе взятых (см. табл. 1), что было неоднократно подтверждено в дальнейших исследованиях (см., однако, результаты из [Воскобойникова 2008]). Отметим при этом, что с точки зрения трех сформулированных на английском материале гипотез этот тип предложений должен был быть, наоборот, наиболее простым.

Таким образом, факт большей синтаксической сложности предложений типа 3 для русскоязычного носителя на сегодняшний день не вызывает сомнений. Однако перед тем как проводить эксперимент, мы предварительно проанализировали Национальный корпус русского языка (www.ruscorpora.ru) на предмет соотношения предложений разных типов (см. табл. 1): как и ожидалось, самые сложные для понимания испытуемых предложения оказались самыми низкочастотными.

2.2. Эксперимент Н. М. Воскобойниковой (2008)

Эксперимент, описываемый в работе [Воскобойникова 2008], был проведен по той же методике разыгрывания сцен, но половина предложений была записана с паузой хезитации внутри именной группы (пример см. ниже). Данный эксперимент имел очень сложный экспериментальный дизайн, поэтому отметим только несколько актуальных для нашей работы результатов: (i) независимо от наличия хезитационной паузы инструкции с прямым порядком упоминания событий понимались значимо хуже инструкций с обратным порядком; в случае отсутствия хезитации они также требовали больше времени на обработку (**анти-гипотеза 1**); (ii) испытуемые, которые сделали мало ошибок (то есть предположительно имели большой объем рабочей памяти), при наличии паузы тратили больше времени на выполнение инструкций с обратным порядком и меньше — инструкций с прямым порядком. Другими словами, при наличии паузы хезитации более легкие предложения понимались испытуемыми с большим трудом, а более трудные — с меньшим.

2.3. Методологический экскурс

Современная психолингвистика исходит из следующих двух интуитивно логичных постулатов: чем человеку сложнее, тем он (i) делает больше ошибок и (ii) тратит больше времени на выполнение задания. Из этих постулатов вытекают два типа методик, уже упоминавшихся ранее — в онлайновых методиках обязательно измеряется время реакции испытуемых, а в оффлайновых — количество допущенных ими ошибок. Однако в том

Таблица 1. Стимульный материал и результаты

Тип	Схема предложения	Частотность по корпусу (из 100%)	ошибки, в % [Федорова 2005]	ошибки, в % [Воскобойникова 2008]
1	Перед тем как Б, А	13	3,8	4,2
2	После того как А, Б	41	2,5	24,3
3	А, перед тем как Б	5	72	14,1
4	Б, после того как А	41	4,4	7,0
	в среднем		20,7	12,4

Таблица 2. Результаты эксперимента

	Тип конструкции	ошибки, %	Среднее время реакции, мс
1	А, перед тем как Б	31	737
2	А, ммм… перед тем как Б	27	609
3	А, перед тем как ммм… Б	31	751
4	Б, после того как А	6,25	790
5	Б, ммм… после того как А	6,25	852
6	Б, после того как ммм… А	5,2	777

случае, когда испытуемый делает очень много ошибок (больше 10%) и/или тратит очень много времени на выполнение задания, мы должны задуматься об экологической валидности нашего эксперимента — если мы даем испытуемому слишком большую нагрузку на его когнитивный аппарат, то его языковое поведение может сильно отличаться от обычного. В эксперименте, описанном в [Федорова 2005], для трех из четырех случаев инструкции были на допустимом по сложности уровне, однако для случая (3) было зафиксировано 72% ошибок (см. табл. 1). В работе [Воскобойникова 2008] инструкции типа *А теперь, перед тем как положить* ***коричневую эээ свинью*** *над топором, положите самолет справа от жука* дали более ровные результаты, но все равно уровень сложности был выше оптимального. Таким образом, в настоящем исследовании мы хотим уменьшить сложность инструкций и тем самым увеличить его экологическую валидность.

С другой стороны, в [Воскобойникова 2008] было совмещено два метода — автор одновременно считал как допущенные ошибки, так и время реакции, что, на наш взгляд, могло дать большое количество «шума»: испытуемый мог дольше думать и ошибаться, потому что ему сложно, а мог затратить мало времени и ошибиться вследствие этого. Для получения более чистых результатов мы сначала отдельно опишем количество и типы ошибок, а затем будет рассматривать время реакции только для тех случаев, когда испытуемый выполнял задания правильно.

2.4. Экспериментальное исследование с носителями русского языка

Наш эксперимент был проведен по аналогичной методике, однако для получения более однозначно интерпретируемых результатов мы значительно упростили его дизайн, оставив только два типа предложений, максимально сложный (тип 3) и его более простой аналог (тип 4), см. табл. 2.

В ходе эксперимента 44 испытуемых передвигали виртуальные объекты в соответствии с инструкциями типа *Положите медведя в правый верхний угол, (ммм) после того как подвинете самолетик на одну клетку вправо*. Однако, вопреки нашим ожиданиям, данный эксперимент не оказался проще: в среднем испытуемые совершили 18%[2] ошибок (надо учитывать, однако, что мы оставили предложения максимально сложного для понимания типа), хотя распределение ошибок оказалось стандартным — их было в пять раз больше в случае конструкции типа 3. Таким образом, в дальнейшем нам необходимо будет еще несколько уменьшить нагрузку на когнитивный аппарат испытуемых. Кроме того, мы не увидели и значимого влияния хезитационной паузы на количество ошибок.

Теперь опишем онлайновую составляющую эксперимента. Во-первых, пауза хезитации влияет на процесс обработки высказывания только в том случае, когда она находится перед союзным словом (примеры 2 и 5); пауза после союзного слова (3 и 6) не оказывает существенного влияния на понимание, не отличаясь от случая без паузы (1 и 4). Данный результат хорошо согласуется с уже имеющимися данными о том, что заполненные паузы хезитации значительно чаще встречаются в начале дискурсивного фрагмента, а не в его середине, [Ferreira & Bailey, 2004]. Во-вторых, оказалось, что в среднем испытуемый тратит наименьшее время на обработку сложных предложений (типа 3), когда перед союзным словом есть пауза хезитации, и наибольшее — на обработку изначально более легких (типа 4), но также при наличии перед союзным словом паузы. Как можно заметить, этот результат перекликается с результатом, полученным Н. Воскобойниковой для испытуемых с небольшим количеством ошибок. На наш взгляд, он может быть объяснен психологическим феноменом **аккомодации внимания**, согласно которому продуктивность выполнения задачи снижается, если время ее выполнения отклоняется от оптимального. Полученные данные также хорошо согласуются с предположением о том, что пауза хезитации сигнализирует, что за ней последует нечто, чего слушающий не ожидает, и он должен приготовиться потратить больше усилий на построение верной

[2] Данные по ошибкам даются по 24 случайно выбранным испытуемым.

ситуационной модели. Если пауза употребляется по этим правилам, слушающему требуется меньше времени на то, чтобы понять синтаксическую структуру предложения, но если пауза употребляется не по тем правилам, по которым слушающий привык ее декодировать (появляется перед ожидаемым элементом), ее появление, наоборот, вызывает дополнительную нагрузку на когнитивный аппарат и замедляет обработку информации. Поэтому изначально более простые предложения с паузой хезитации требуют больше времени по сравнению со средним уровнем (когда пауза отсутствует), а более сложные — меньше времени.

Заключение

Итак, в эксперименте с носителями русского языка хезитационная пауза, сделанная перед началом артикуляции дискурсивного фрагмента, действительно является маркером возникающих у говорящего сложностей, что, в свою очередь, оказывает влияние на слушающего, который настраивается на обработку более сложной конструкции. Если конструкция на самом деле оказывается сложной, времени требуется гораздо меньше, но если, против ожидания, простой — испытуемому требуется дополнительное время на то, чтобы перестроиться.

Литература

1. *Воскобойникова Н. М.* Роль хезитации в понимании сложноподчиненных предложений с придаточными времени. Дипломная работа, МГУ, 2008.
2. *Подлесская В. И. & Кибрик А. А.* Речевые сбои и затруднения // А. А. Кибрик, В. И. Подлесская (ред.) Рассказы о сновидениях: Корпусное исследование устного русского дискурса. М.: ЯСК, 2009. С. 177–218.
3. *Николаева Т. М.* Новое направление в изучении спонтанной речи (о так называемых речевых колебаниях) // Вопросы языкознания, № 3, 1970.
4. *Федорова О. В.* Перед или после: что проще? (понимание сложноподчиненных предложений с придаточными времени) // Вопросы языкознания, № 6, 2005.
5. *Шмелев А. Д.* Показатели хезитации в устной русской речи // Язык. Личность. Текст. Сб. ст. к 70-летию Т. М. Николаевой. М.: ЯСК, 2005. С. 518–529.
6. *Arnold J. E., Fagnano M. & Tanenhaus M. K.* Disfluencies signal theee, um, new information // Journal of Psycholinguistic Research, 32(1), 2003.
7. *Clark H. H. & Clark E. V.* Semantic distinctions and memory for complex sentences. // Quarterly Journal of Experimental Psychology, 20, 1968.
8. *Clark H. H. & Wasow T.* Repeating words in spontaneous speech // Cognitive Psychology, 37, 1998.
9. *Clark H. H. & Fox Tree J. E.* Using uh and um in spontaneous speaking // Cognition, 84, 2002.
10. *Corley M., MacGregor L. J. & Donaldson D. I.* It's the way that you, er, say it: Hesitations in speech affect language comprehension. Cognition, 105, 2007.
11. *Corley M. & Stewart O. W.* Hesitation disfluencies in spontaneous speech: The meaning of um // Language and Linguistics Compass, 4, 2008.
12. *Ferreira F.* Effects of length and syntactic complexity on initiation times for prepared utterances // Journal of Memory and Language, 30, 1991.
13. *Ferreira F. & Bailey K. G. D.* Disfluencies and human language comprehension // Trends in Cognitive Sciences, 8–5, 2004.
14. *Fox B. A., Hayashi M. & Jasperson R.* Resources and repair: a cross-linguistic study of syntax and repair // Interaction and grammar. Cambridge: Cambridge University Press, 1996.
15. *MacGregor L. J., Corley M. & Donaldson D. I.* Not all disfluencies are are equal: The effects of disfluent repetitions on language comprehension // Brain & Language, 111, 2009.
16. *O'Connell D. C. & Kowal S.* The history of research on the filled pause as evidence of the written language bias in linguistics (Linell, 1982) // Journal of Psycholinguistic Research, Vol. 33, № 6, 2004.

Пересказывательность в русском языке[1]

Quotation and rendering markers in russian

Левонтина И. Б. (irina.levontina@mail.ru)

ИРЯ им. В. В. Виноградова РАН, Москва

О пересказывательности обычно говорят в контексте категории эвиденциальности (засвидетельствованности), которая существует в качестве грамматической категории (наклонения или подобной) в некоторых языках — в частности, индейских и тибето-бирманских, болгарском, литовском, турецком. Разумеется, подобные значения выражаются и в других языках. Применительно к русскому языку в этой связи обычно упоминаются так называемые ксенопоказатели — частицы *мол, дескать, де*, а также *якобы* и *грит (гыт)*. Однако оказывается, что арсенал средств, служащих в русском языке для оформления цитирования или пересказа, гораздо богаче и представлены они на разных уровнях. В докладе рассмотрены некоторые лексические и просодические маркеры пересказывательности.

О пересказывательности обычно говорят в контексте категории эвиденциальности (засвидетельствованности), которая существует в качестве грамматической категории (наклонения или подобной) в некоторых языках — в частности, индейских и тибето-бирманских, болгарском, литовском, турецком. См., например, [Slobin, Aksu 1982; Chafe, Nichols 1986; Anderson 1986; Willett 1988; Bybee et al. 1994; De Haan 1998; Эвиденциальность 2007].

Так, Р. О. Якобсон в классической работе о шифтерах упоминает пересказывательное наклонение в болгарском. Он рассматривает диалог, в котором человек отвечает на вопрос о том, что произошло с лодкой: он «…сначала ответил заминала ‘говорят, что отплыла’, а потом добавил: замина ‘я свидетель, что отплыла’» [Якобсон 1972, 101].

Разумеется, подобные значения выражаются и в других языках. Можно упомянуть хотя бы использование формы первого конъюнктива для выражения этого смысла в немецком; ср. *Er habe das vergessen* (‘Он, по его словам, забыл’), в отличие от фразы с индикативом *Er hat das vergessen* (‘Он забыл’)[2].

Применительно к русскому языку в этой связи обычно упоминаются так называемые ксенопоказатели — частицы *мол, дескать, де*, а также *якобы* и *грит (гыт)*. Этимологически они в основном связаны с глаголами говорения. Особенно много обсуждались *дескать* и *мол* [Отин 1966; Колодезнев 1969; Fontain 1983; Камю 1992; Баранов 1994; Арутюнова 2000; Шестухина 2003]. Эти две частицы используются и в прямой, и в косвенной, и в несобственно-прямой речи, будучи при этом позиционно достаточно свободными. Функция их состоит в том, чтобы, по словам Н. Д. Арутюновой, «маркировать присутствие Другого» [Арутюнова 2000: 448]. При этом как чужая понимается и собственная речь, сказанная ранее или планируемая на будущее, и интерпретация поведения другого человека, его реакции и т. д. Поскольку говорящий при помощи ксенопоказателей отстраняется от позиции другого человека, прагматически такие слова притягивают разного рода оценки, чаще отрицательные, воспроизводимой речи.

В работе [Камю 1992] сделана попытка установить семантические различия между частицами *мол* и *дескать*. Различия это слабые и имеющие характер предпочтений. Суть их сводится к тому, что *мол* более тесно связано с исходным высказыванием, тогда как *дескать* допускает более вольную трактовку ситуации.

Наиболее тонко различия между *мол, дескать*, а также *де* описаны в статье [Баранов 1994]. Значение данных частиц рассмотрено здесь в контексте оппозиции «свой-чужой», в частности через идею коммуникативной ответственности: «Рассматривая семантику *мол* и *дескать* в рамках идеи коммуни-

[1] Работа выполнена при поддержке гранта Президента Российской Федерации для государственной поддержки ведущих научных школ Российской Федерации НШ-4019.2010.6.

[2] Мы не говорим, естественно, о том, что этот смысл всегда можно выразить описательно (*Он сказал, что…, По его словам…*). В письменной речи цитирование оформляется кавычками. В устной публичной речи аналогами кавычек являются формулы «Цитата» и «Конец цитаты». Ср. также распространившийся в последние годы заимствованный лекторский иконический жест «кавычки», выполняемый указательнымисреднимпальцамиобеихрукодновременно.

кативной ответственности, можно утверждать, что *дескать* отражает нежелание говорящего брать на себя ответственность за чужое (в целом или частично), а мол, напротив, свидетельствует о том, что за какие-то фрагменты чужого опыта он готов разделить ответственность с автором цитаты.

При таком подходе существенным оказывается выявление сферы действия своего и чужого и, соответственно, сферы разделения ответственности и отказа от нее»[3] [Баранов 1994: 116].

Возвращаясь к репертуару ксенопоказателей в русском языке, можно отметить, что слово *якобы* имеет очевидные и гораздо более существенные отличия от *мол* и *дескать*. *Якобы* несовместимо с прямой речью, оно связано с передачей чужой речи скорее de re, а не de dicto. Кроме того, *якобы* выражает сомнение в достоверности передаваемого сообщения.

Однако оказывается, что арсенал средств, служащих в русском языке для оформления цитирования или пересказа, гораздо богаче. Так, соответствующее значение есть у слова *ах*; ср:

— Время идет быстро, а между тем здесь такая скука! — сказала она, не глядя на него. — Это только принято говорить, что здесь скучно. Обыватель живет у себя где-нибудь в Белеве или Жиздре — и ему не скучно, а приедет сюда: «Ах, скучно! ах, пыль!» Подумаешь, что он из Гренады приехал. (А. П. Чехов, Дама с собачкой).

Здесь очень важно, что вне контекста пересказывания *ах* в подобной фразе употребить нельзя; ср.:

*— Вы хорошо съездили? — * Ах, скучно! ах, пыль!*

Междометие *ах* в своем основном значении соответствует определенной части эмоционального спектра (ср. цветаевское «*Ах», когда чудно*). Однако в нашем случае ограничения на тип эмоции если не снимаются, то по крайней мере ослабляются. *Ах* в рассматриваемом значении может интонироваться двумя способами. Либо оно выступает в качестве проклитики (*ахскучно, ахпыыль*. Так интонирует эту чеховскую фразу знаменитый чтец Дмитрий Журавлев), при этом слово растягивается и произносится с пологим повышением тона с последующим падением. Либо *ах* произносится отдельно, тогда повышение тона делается на нем, а падение на следующем слове (Так произносит другой исполнитель «Дамы с собачкой» Игорь Ясулович).

Приведем еще несколько примеров:

— Рак. (Параджанову недавно сделали операцию после того, как у него обнаружили рак лёгкого.) Начинает причитать: — Ах, я скоро умру! Посмотри, какой у меня шрам тут. Вообще я сейчас пойду лягу и буду умирать. Сначала только чаю попьём. [Сати Спивакова. Не всё (2002)]

Заметим, что само по себе восклицание *Ах, я скоро умру!* Представить себе трудно.

Только что перечитал этот кусок и подумал, что вышло как-то уж очень сумбурно. Ну что это значит — «мелькнуло лицо… Мышкина»? Надо было как-то по-другому. Рассказать, как он вошёл, как представился, что сказал… А то: ах, извините! Ах, у меня память отшибло! Хотя… Беда-то в том, что её у меня и правда отшибло. Не помню я, как он у нас очутился, не помню, с чего начал… [Вера Белоусова. Второй выстрел (2000)]

Здесь интересно то, что в действительности в предшествующем тексте не было ни *извините*, ни *память отшибло*. Это типичная передача de re, причем очень вольная.

Все эти книжонки, монастыри, путешествия по «святым местам» на собственных «Волгах» сделались модой и оттого пошлостью. Раньше все скопом на Рижское взморье валили, а нынче — по монастырям. Ах, иконостас! Ах, какой нам дед встретился в одной деревеньке! А самовары? Иконы? [Юрий Трифонов. Предварительные итоги (1970)]

Данный пример интересен тем, что слово *ах* вполне пригодно для выражения восторга. Поэтому этот контекст можно интерпретировать двояко: либо в том смысле, что другие люди так и говорили *Ах, иконостас!* и т. д., либо (скорее), что *ах* появляется именно в пересказе.

Аналогично устроена и фраза из «Горя от ума»: *Ах, Франция, нет в мире лучше края, решили две княжны…* Княжны и сами могли сказать *Ах, Франция!,* но возможно, это пересказ фразы типа *Франция — лучшее место на свете*.

К уточнению значения *ах* мы вернемся чуть позже, а пока обратимся к еще одному ксенопоказателю. В работе [Подлесская, Кибрик 2009] отмечается специфическое употребление слова *вот* как средства «передачи угрозы и осуждения в чужой речи»; ср.:

\тоже на меня ’ /посмотрела ,
«/\Вот!
Я тебя /–в-выгоню-у и-из-зз ··(0.2) этой из ш=
··(0.2) ’из ··(0.2) ’ /школы!»,

«Изысканность этого употребления состоит в том, что цитирование происходит в форме прямой

[3] Имеется в виду план выражения, форма выражения смысла, сам смысл — в духе концепции Е. Гофмана о расщеплении говорящего на автора идеи, автора способа оформления мысли и субъекта говорения, которую Баранов дополняет еще одной ролью говорящего — ролью «коммуникативного стратега» [Баранов 1994: 116].

речи, однако сам маркер не мог принадлежать исходной реплике — при непосредственном выражении угрозы в диалогическом режиме вот употреблено быть не может. Таким образом, маркер вот, подобно более нейтральным маркерам чужой речи типа мол, может служить средством перевода прямой речи в «несобственно-прямую». Заметим при этом, что мол — в отличие от вот — не способно к автономизации, эта частица никогда не акцентируется и всегда встроена в пропозициональную ЭДЕ».

Использование слова *вот* как ксенопоказателя чрезвычайно частотно в устной речи, но очень плохо фиксируется в речи письменной, поскольку для реализации этого значения необходим особый просодический контур.

Надо заметить, что *вот* передает далеко не только угрозу. Ср.:

А она сидит и ноет: «Вот, я такая несчастная…»;

Он расхвастался: «Вот, я самый крутой»;

Привязалась: «Вот, как тебе не стыдно, что у тебя за юбка»;

А он все обещает: «Вот, деньги будут со дня на день, все отдам».

Существенно, что говорящий даже не обязательно осуждает речевую деятельность другого человека, которую он воспроизводит. Ср.

Ну и что же, то она первая позвонила? А ты бы ей сказал: «Вот, я сам собирался тебе позвонить, поздравить».

Теперь вернемся к слову *ах* и рассмотрим сходства и различия между двумя ксенопоказателями — *ах* и *вот*.

Обе единицы используются как бы в качестве открывающей кавычки, маркируя начало чужой речи. При этом, в отличие от частиц типа *мол*, эти две единицы не предполагают, что за ними последует подробное изложение или тем более точное воспроизведение чужой речи. Обе они вводят обычно сокращенный пересказ, однако *вот* скорее склонно к тому, чтобы передать одну или несколько наиболее важных реплик, а *ах* скорее передает общий смысл речи и ее эмоциональный настрой. При этом *ах* обычно подразумевает, что человек, речь которого передается, выражал какие-то оценки или был эмоционален (возможно, чрезмерно).

В работе [Подлесская, Кибрик 2009] отмечается, что «средством передачи угрозы и осуждения в чужой речи является даже не сам маркер *вот*, а реализуемая на нем особая просодическая фигура с восходяще-нисходящим тоном. Эта же фигура, с тем же значением, может реализоваться и на ряде других маркеров, например, на частице *а*, тоже способной к автономному употреблению». Приводится следующий пример:

И \каждое н-на \меня=:
«/\'А-а!
/\Вот!
Ты /\двоечница!
\С-смотри на своих оценки¡!»

В книге [Янко 2008: 109] описывается интонационный контур, названный «интонацией ментальной деятельности, т. к. этот интонационный тип направлен на отражение средствами интонации разного рода информационных и мыслительных процессов». В этой связи упоминаются припоминание, недоумение, погружение в мечты, а также и передача чужой речи (*Тетя сказала, надо чего-то там уколы делать*). Эта интонация описывается так: «Соответствующий акцент характеризуется подъемом тона и существенным удлинением ударного слога акцентоносителя ремы. Вся заударная область ровная (иногда с небольшим естественным падением)».

В этой работе справедливо отмечается, что «при передаче чужой речи говорящий не копирует интонацию того, чья речь цитируется, а придает соответствующему фрагменту речи просодию воспоминаний».

Как кажется, просодическое оформление чужой речи нуждается в гораздо более детальном описании. Ограничимся пока некоторыми предварительными замечаниями.

При сходстве интонационного контура пересказывания и, например, припоминания, между ними есть существенные различия. Во-первых, при пересказывании возможен эмоциональный вариант, при котором фраза интонируется гораздо более эмфатически, и такое произношение невозможно ни при припоминании, ни при выражении недоумения или мечты.

Во-вторых, для пересказывания очень характерно дробление чужой речи на более мелкие сегменты, чем это естественно для обычной речи, даже на отдельные слова; ср.:

— И что он ответил?
— Да что ответил! «/\Маама не разре/\шаает»[4].

В пересказе чужая речь часто предстает как бы ритмизованной, произносимой с периодическим подъемом с последующим падением тона на ударных слогах (это чем-то напоминает перечислительную интонацию); ср.

[4] Мы пока используем здесь значок /\, сознавая, однако, что возможные здесь интонационные контуры нуждаются в специальном изучении, в том числе инструментальными средствами.

А он мне и говорит: «/\Воот, /\деевушка, какая вы кра/\сиивая, а пой/\деемте, погу/\ляем, а /\дайте теле/\фоончик».

Надо заметить, что разложение чужой речи на бессвязные фрагменты характерно не только для просодии, но и для синтаксиса пересказывания. Это отмечается в работе [Камю 1992].

Интонация пересказывания так искажает исходную фразу, что слушающему сразу понятно, что это не собственные слова говорящего. Например, в вопросе *Как вас зовут?* повышение тона на *зовут* в литературном языке невозможно. А в цитате оно естественно (конечно, не по ИК-3, а более пологое, ближе к ИК-6, с некоторым понижением к концу ударного слога).

Наконец, если говорящий негативно оценивает пересказываемую речь (а при введении в текст чужой речи она очень часто так или иначе оценивается), может возникать явление передразнивания. Существуют разные фонетические и даже мимические средства передразнивания. Отметим хотя бы два. В той же работе [Камю 1992] со ссылкой на устное сообщение С. В. Кодзасова обращается внимание на «частое явление назализации, сопровождающее значение передразнивания».

Другое фонетическое средство передразнивания — это «блеяние» (*ма-а-ама*).

Разумеется, список способов передразнивания может быть продолжен.

Есть еще одно явление, связанное с приемами передачи чужой речи, которое необходимо отметить. Чужая речь может передаваться не путем собственно пересказывания, а путем имитации речи. Кроме ксенопоказателей, существуют и заместители речи — бессмысленные сочетания, обычно содержащие повторы и рифмы, на которых часто реализуется тот же или подобный интонационный контур, который был рассмотрен выше. Это сочетания типа *ля-ля тополя, ляля-фа-фа, тэ-тэ-тэ, тэ-тэ-нэ-нэ, тэто-это*, а также довольно новое заимствование *бла-бла(-бла)*;

Я ему объясняю: «У меня много работы, а завтра теща приезжает, тэ-тэ-нэ-нэ <ля-ля тополя>…»;

Ты ей скажи, что ты к нему хорошо относишься, но только как к другу, бла-бла-бла.

Ср. также:

Прибегает: «А! О!» А чем я могу ему помочь?;

Опять наехала на меня: «Аа! Даа!» Надоела уже.

Сюда можно добавить и единицы *тыры-пыры* и *тыр-пыр восемь дыр*.[5] Они практически не отражаются в письменных текстах, но в устной речи довольно часто используются, особенно если перед этим уже был какой-то намек на содержание чужой речи и продолжение, в общем-то, понятно. Некоторые из этих выражений используются не только для замещения, но и для обозначения чужой речи; ср.

Вот сейчас, я пролистал ЖЖ и думаю что-то вроде — вот я лошара и неудачник, так бездарно проводил время и сох па разным тёлкам и всё бес толку и писал какую-то пургу какие-то рассказы и бла бла.

Стоит заметить, что зачастую функцию маркеров пересказывательности выполняют единицы с более широким значением. Так, в этой роли часто выступает слово *вроде* — показатель неточного знания. Ср. *Ты вроде похудела* [кажется, говорящий не уверен в своей оценке] и *Он вроде уволился* [говорят][6]. Так же может использоваться и показатель приблизительности номинации *типа*: *А она стала говорить, что муж типа так занят*. Интересно также недавно пришедшее в язык словечко *такой* в изобразительной функции (пожалуй, по значению оно ближе всего к настоящему изобразительному времени). Иногда его считают показателем эвиденциальности, поскольку оно в большинстве случаев вводит прямую речь; ср. А *я такая: «Как тебя зовут?»* [Савчук, в печати]. Это, однако, не обязательно; ср. *А я такая подхожу, беру сигарету и закуриваю. Все в шоке*.

Возвращаясь к интонации пересказывания, отметим следующее. Она свободно сочетается с разными ксенопоказателями, однако сама имеет более широкое значение. В частности, тот же интонационный контур регулярно используется для оформления тривиального содержания. Так, в научной речи докладчик часто вынужден в начале доклада, прежде чем перейти к сути дела, излагать сведения, которые кажутся ему банальными и общеизвестными, но без которых он не может обойтись в изложении своего сюжета. В этом случае он часто пользуется интонационным контуром, характерным для цитирования. Это совершенно понятно: тривиальное подается как уже сказанное другими и соответственно интонируется. Частицы же больше связаны с конкретным высказыванием другого человека и потому в подобных контекстах едва ли возможны.

Итак, мы увидели, что средства выражения пересказывательности представлены в русском языке на разных уровнях. Ни в коей мере не претендуя на полноту описания, мы кратко рассмотрели некоторые из них. Наверняка существуют и другие. Таким образом, функционирование категории пересказывательности в русском языке нуждается в дальнейшем изучении.

---

[5] Последние напомнил мне один из анонимных рецензентов «Диалога», за что ему (ей) отдельное спасибо.

[6] На подобные контексты обратил мое внимание А. Д. Шмелев.

Литература

1. *Арутюнова Н. Д.* Показатели чужой речи *де, дескать, мол* //Язык о языке. Под общ. рук. и ред. Н. Д. Арутюновой. — М., 2000. С. 437–452.
2. *Баранов А. Н.* Заметки о дескать и мол // ВЯ 1994, № 4, с. 114–124.
3. *Камю Р. Мол, дескать, -де*: чужая речь в чужом языке // Проблемы интенсивного обучения неродным языкам (материалы первой международной научно-методической конференции, 27–28 мая 1992 г.), Российский государственный педагогический университет имени А. И. Герцена, изд. «Образование», Санкт-Петербург, 1992, 52–57.
4. *Колодезнев В. М.* О значении частиц *мол, де, дескать* // Русский язык в школе 1969, № 1.
5. *Отин Е. С.* О субъективных формах передачи чужой речи // Русский язык в школе 1966, № 1.
6. *Подлесская В. И., Кибрик А. А.* Дискурсивные маркеры в структуре устного рассказа: опыт корпусного исследования // Диалог 2009.
7. *Савчук С. О.* Местоимение такой в функции маркера чужой речи в устном высказывании // «Вопросы культуры речи», вып. 10 (в печати).
8. *Шестухина И. Ю.* Эмоционально-экспрессивное и функционально-стилистическое значение частиц *мол, де, дескать* в русском языке // Мат. Всероссийской научно-практ. конф. «Русский язык и культура речи как дисциплина государственных образовательных стандартов высшего профессионального образования: опыт, проблемы, перспективы». — Барнаул, Изд-во АГУ, 2003. — С. 366–368.
9. *Эвиденциальность* в языках Европы и Азии. Сборник статей памяти Наталии Андреевны Козинцевой.. М., Наука, 2007
10. *Якобсон Р. О.* Шифтеры, глагольные категории и русский глагол / Принципы типологического анализа языков различного строя. М., 1972.
11. *Янко Т. Е.* Интонационные стратегии русской речи в сопоставительном аспекте. М., 2008.
12. *Fontain J.* Grammaire du texte et aspect du verbe russe contemporain. P., 1983.
13. *Anderson L.* (1986). «Evidentials, paths of change, and mental maps: typologically regular asymmetries.» In Chafe & Nichols (eds.). (1986), 273–312.
14. *Bybee J., Perkins R. and Pagliuca W.* (1994). The evolution of grammar: Tense, aspect and modality in the lanuages of the world. Chicago: University of Chicago Press.
15. *Chafe W. and J. Nichols* (eds.) (1986). Evidentiality: the linguistic coding of epistemology. Norwood: Ablex, 1986.
16. *Ferdinand H., De* (1998). The Category of Evidentiality. Ms.
17. *Slobin D. and Aksu A.* (1982). «Tense, Aspect, and Modality in the use of the Turkish evidential». In Hopper P. J. (ed.) Tense-aspect: Between Semantics and Pragmatics. Amsterdam: Benjamins, 185–200.
18. *Willett T.* (1988). «A cross-linguistic survey of grammaticization of evidentiality». Studies in Language 12.1, 57–91.

Опущение прямого объекта и близкие процессы в арабском языке в сопоставление с русским (на материале лингвистических корпусов)[1]

Direct object omission and similar processes in arabic in comparison with russian (based on corpus data)

Летучий А. Б. (alexander.letuchiy@gmail.com)

Институт русского языка РАН, Москва

В докладе на корпусном материале рассматриваются опущение прямого объекта и варьирование управления глаголов в арабском языке в сопоставлении с русским. Выявляются различия между языками: в арабском языке в меньшей мере, чем в русском, распространено опущение референтного определённого объекта, а опускаются, в основном, неопределённые нереферентные объекты.

Введение

В синтаксическом описании языка большое место занимает понятие переходности. Глаголы делятся на переходные (способные принимать прямой объект, в русском языке — дополнение в винительном падеже) и непереходные (неспособные его принимать).

Однако это важное понятие является в то же время одним из самых проблемных. Дело в том, что деление глаголов на переходные и непереходные отнюдь не однозначно.

Во-первых, существуют лабильные глаголы, способные употребляться и как переходные, и как непереходные. В работах [Недялков 1969], [Haspelmath 1993], [Kazenin 1993], [Летучий 2006] описано распределение лабильности по семантическим группам глаголов в языках мира, здесь она рассматриваться не будет. Во-вторых, даже у строго переходных глаголов объект может быть не выражен. Например, русский глагол *удивить* нередко употребляется в безобъектной конструкции:

(18) Правда, несколько удивляет то, что значительная часть комментариев сводилось к запрету на так называемое сквернословие (сквернословие матерную брань). [Алексей Шмелев. Вопросы языкознания в Государственной думе // «Отечественные записки», 2003]

Такого рода процессы редко адекватно описываются в грамматиках, словарях и даже более специальных синтаксических работах (см., впрочем, работу [McShane 2005], где предложена классификация случаев эллипсиса актантов глагола). Причиной является то, что опущение требует наблюдения над статистикой употребления глаголов с объектом и без объекта в текстах.

Предметом нашего описания является опущение прямого объекта (далее для краткости мы иногда будем употреблять просто слово «объект») в арабском языке в сопоставлении с русским. Мы хотим показать, что разные типы опущения в разной мере распространены в разных языках, и выявить распределение опущений.

Опущение и близкие процессы исследовались нами на корпусном материале: для русского языка использовался Национальный корпус русского языка (www.ruscorpora.ru), для арабского — корпус ArabiCorpus (http://arabicorpus.byu.edu/).

1. Некоторые синтаксические свойства арабского языка

Арабский язык имеет всего три падежа: именительный, родительный и винительный. Приглаголь-

[1] Исследование выполнено в рамках гранта РФФИ 10-06-00338 и гранта президента РФ для государственной поддержки молодых кандидатов наук МК-3522.2010.6.

ное дополнение может кодироваться либо винительным падежом, либо предложной группой. Прямые объекты всегда имеют форму винительного падежа.

В арабском имеется большой набор словообразовательных моделей («пород»), изменяющих набор и свойства актантов глагола, а также его аспектуальную характеристику: ср., например, *'anha:* 'закончить', 4 порода — *intaha:* 'закончиться', 8 порода (подробнее о породах см. [Гранде 1998]).

2. Общие закономерности

В целом по корпусу количество структур с опущением объекта крайне мало. Впрочем, по-видимому, это вообще естественно для письменной речи.

Так, в корпусе, включающем египетские пьесы, стиль которых имитирует разговорный, среди 50 первых примеров встречается всего 1 пример с опущением объекта. Даже в этом случае опущение сомнительно: объект не выражен при переходном глаголе *kharaǧa* 'выходить', для которого естественно употребляться безобъектно: известно, что глаголы с такой семантикой зачастую употребляются без указания исходной точки, если она понятна из контекста или из знания ситуации.

В корпусе газет мы наблюдаем ту же картину: среди первых примеров только в одном примере переходный глагол выступает без прямого объекта — и даже в этом примере объект не опускается, а кодируется с помощью предложной группы.

Наконец, количество примеров с интересующими нас явлениями в художественной литературе несколько больше, но и здесь мы скорее имеем дело с вариативностью выражения объекта, а не с его опущением:

(19) *'amsak-a* *bi* *ḥaqi:bat-i-hi*
прикоснуться.IV-3SG.M в сумка.SG-GEN-3SG.M[2]
'Он прикоснулся к своей сумке.'

Глагол *'amsaka* 'прикоснуться' допускает выражение Пациенса либо в винительном падеже, либо с помощью предлога *bi* 'в, с'.

В русском доля примеров с невыраженным объектом несколько выше (8 на 100 примеров). Часть этой, и без того не слишком большой группы, составляют примеры с эллипсисом объекта, выраженного в том же предложении (тип 1 в классификации, предлагаемой в части 3):

(20) Однако при этом заданных полторы недели назад тенденций совершенно не изменил. Напротив — расширил их и углубил. [Андрей Митьков. Беги-стреляй. Итоги второго этапа Кубка мира по биатлону (2002) // «Известия», 2002.12.16]

Объект *их* выражен при первом глаголе *расширить*, поэтому при втором — *углубить* — не выражается.

Другой подкласс случаев, который также не следует отождествлять с «полноценным» опущением объекта, — смена управления глагола:

(21) С небольшой старинной фотографии смотрит девушка с толстой косой, с широкоскулым, широкоглазым и большеротым лицом. [Фазиль Искандер. Слово (1980–1990)]

У глагола *смотреть* есть переходное употребление (*смотреть фильм*) и непереходное (*смотреть на него*). При этом в обоих употреблениях стимул восприятия выражается, хотя и по-разному оформлен.

Тем самым, остаётся очень мало (4) собственно примеров опущения, ср., например, (5):

(22) В месте светлом, в месте злачном, в месте покойном — Так... Начали... Мотор! Хлопнула хлопушка. [Г. Я. Бакланов. В месте светлом, в месте злачном, в месте покойном (1995)]

В данном случае глагол *начать* употребляется без объекта — объект домысливается исходя из ситуации ('начать съёмки').

Далее мы рассмотрим некоторые из конкретных типов опущения и вариативности выражения объекта.

3. Типы опущения

Ориентируясь на русский материал, мы выделяем следующие типы случаев, при которых актант (в нашем случае — прямой объект) может не быть выражен[3]:

1. **синтаксический эллипсис** (опущение одного из вхождений объекта, выраженного в том же предложении, обязательное или допустимое в данном контексте по структурным соображениям, например, *Они увидели его на улице и убили [его]*)

[2] Используемые сокращения: ACC — винительный падеж, GEN — родительный падеж, NOM — именительный падеж, SG — единственное число, PL — множественное число, CONJ — конъюнктив. Римскими цифрами обозначаются номера пород, то есть словообразовательных глагольных моделей. Отметим, что хотя в арабском языке различаются показатели так называемого сопряжённого состояния (существительного с артиклем или зависимым в генитиве) и свободного состояния, для упрощения глоссирования мы это не учитываем.

[3] В нашей классификации мы во многом ориентируемся на положения работы [McShane 2005], однако на сегодняшний день используем менее дробную классификацию.

2. опущение

2.1. опущение нереферентного актанта (*Убивать — грех*)

2.1.1. опущение нереферентного актанта без дополнительных условий (*Он убивал уже не раз*)

2.1.2. опущение нереферентного актанта в контексте модального глагола (*Эти парни умеют убивать*).

2.1.3. опущение нереферентного актанта при неактуальной ситуации

2.2. опущение референтного предупомянутого актанта (*Вася был учителем в школе. Школьники часто утомляли, но вообще работа ему нравилась.*)

2.3. опущение объекта-участника коммуникации (*Как же бесят [меня] эти студенты!*)

3. лабильность (наличие у переходного глагола второго употребления, не предполагающего наличия объекта даже на синтаксическом уровне, например, переходное употребление *Вася льёт воду в кастрюлю* и непереходное *Из труб лила вода*)

Мы не будем обсуждать в данной работе лабильность — явление, тесно связанное с лексической семантикой глагола и предполагающее смену набора *семантических* валентостей, — а также эллипсис, который зачастую мотивирован чисто синтаксически.

В классификацию не попало ещё одно близкое явление, упомянутое выше в связи с примером (5) — смена управления глагола. Ниже мы скажем о ней несколько слов.

Безусловно, предложенную классификацию нельзя считать абсолютно строгой и полной: в частности, могут существовать и другие варианты опущений. При составлении классификации мы ориентировались просто на встреченные нами в Национальном корпусе русского языка случаи опущения объекта.

3.1. Опущение нереферентного объекта

3.1.1. Опущение нереферентного объекта без дополнительных условий

Вне контекста модального глагола нереферентные объекты в русском языке опускаются довольно редко. Так, при глаголах сильного воздействия на одушевлённый Пациенс в русском языке практически нет случаев опущения объекта:

(23) *?Он часто бьёт.*

(24) *?Он часто ломает.*

Мы проанализировали контексты, где финитные формы глагола *убивать*, обладающего сильной семантической транзитивностью, встречаются непосредственно перед точкой (поскольку дополнение в русском языке, как правило, следует после глагола, весьма вероятно, что в таких случаях мы будем наблюдать опущение объекта).

Всего такого рода случаев встретилось 3745[4]. Прежде всего, как уже говорилось, мы исключили из рассмотрения случаи эллипсиса — синтаксического опущения актанта, совпадающего с актантом другого глагола, выраженного в близком синтаксическом контексте (см. выше (3)).

Далее мы рассмотрели первые 200 примеров, где объект не выражен. Помимо случаев эллипсиса, встречаются примеры, где опущение не обусловлено синтаксическим сокращением совпадающей именной группы. Оказалось, что всего в пяти примерах опущение связано с неопределённостью объекта, например:

(25) Эти добрые господа … впускают в нашу простую, проходящую в метро и магазинах жизнь людей, которые уже убивали и насиловали и готовы убивать. [Сергей Есин. Выбранные места из дневника 2001 года (2003) // «Наш современник», 2003.06.15]

В арабском языке опущение наиболее характерно именно для нереферентрных объектов. Так, при глаголах *Daraba* ‘бить, бомбить’, *qatala* ‘убивать’, *haddada* ‘угрожать’ обобщённый объект может не выражаться:

[4] Отметим, что арабские данные менее статистически надёжны, так как базируются на анализе отдельных примеров: статистические данные, аналогичные русским, для арабского корпуса недоступны, поскольку в нём нет поиска по знакам препинания и грамматическим характеристикам. В этой же связи для арабского языка невозможно привести общее количество структур с опущениями прямого объекта при переходных глаголах.

(26) *Huna:ka irha:b-un y-aDrib-u wa laysa dawlat-u-na*
там террорист.PL-NOM 3SG.M-бить-SG но не.быть-3SG.M страна.SG-NOM-1PL
‘Есть террористы, которые бомбят, но это не наша страна.’

(10) *Wa y-aDrib-u wa y-aqtul-u wa y-uhaddid-u fi: jiha:t-i ad-dunya: al-arba‘-at-i*
и 3SG.M-бить-SG и 3SG.M-убивать-SG и 3SG.M-угрожать-SG в сторона.PL-GEN DEF-мир.GEN.SG DEF-четыре-F-GEN.SG
‘Он бомбит, убивает и угрожает (всем) во всех четырёх сторонах света’.

Такое опущение возможно при данном глаголе и в русском языке, хотя и выглядит несколько странно (ср. *?Они постоянно бомбят*). Для глагола *угрожать* в русском языке вообще невозможно нереферентное опущение объекта: ср. примеры, где опускается только референтный определённый объект:

(11) Маршалл медлил, не видя военной необходимости. Гровс настаивал, угрожал и всё же добился. [Даниил Гранин. Зубр (1987)]
(12) Левий Матвей говорил, что не хочет расстаться с этим телом. Он был возбуждён, выкрикивал что-то бессвязное, то просил, то угрожал и проклинал… [М. А. Булгаков. Мастер и Маргарита, часть 2 (1929–1940)]

Однако в то же время референтное опущение вида *Но ты же ударил первым!*, судя по всему, не слишком частотно в арабском языке — во всяком случае, в выборке примеров на глаголы *Daraba* ‘бить, бомбить’, *qatala* ‘убивать’ такие случаи опущения отсутствуют.

3.1.2. Опущение нереферентного объекта при модальных глаголах

Одним из факторов, облегчающих опущение нереферентного объекта, являются модальные глаголы (исследовались вхождения модального глагола *ʾistaTaʿa* ‘мочь’). Так, в сочетании с глаголом *qatala* ‘уби(ва)ть’ с модальным глаголом его объект опускается в 3 из 6 случаев.

(13) *y-astaTi:ʿ-u ʾan y-aqtul-a da:khila al-kuwayt-i aw khariğa-ha:*
3SG.M-мочь-SG чтобы 3SG.M-убить-CONJ внутри DEF-Кувейт-GEN.SG или вне-3SG.F
‘Он может убивать в Кувейте или за его пределами.’

В отсутствие модального глагола опущения при *qatala* сравнительно редки: среди первых 50 примеров в поисковом запросе всего в 8 наблюдается опущение объекта.

Точно так же при глаголе *kaddaba* ‘врать, обманывать’ при наличии модального глагола может выбираться непереходная модель без дополнения:

(14) *Wa huwa y-astaTi:ʿ-u ʾan y-ukaddib-a*
и он.NOM 3SG.M-мочь-SG чтобы 3SG.M-врать-SG
‘А он может врать.’

В то же время эта тенденция соблюдается не для всех глаголов. Так, для глагола *raʾa* ‘видеть’ даже при модальном глаголе найден всего один пример с опущением:

(15) *la y-astaTiʿ-u ʾan yara:*
не 3SG.M-мочь-SG чтобы 3SG.M-видеть.SG.CONJ
‘Он не может видеть (не видит, незрячий).’

Точно так же прямой зависимости не наблюдается при глаголе *ʿarafa* ‘знать’: структуры с опущением при нём довольно редки, хотя и встречаются:

(16) *Ad-dawr-u mula:ʾim-u la-hu li ʾanna-hu ma y-aʿrif-u y-aḥki:*
DEF-роль.SG-NOM мучительный для-3SG.M для что-3SG.M NEG 3SG.M-знать-SG 3SG.M-рассказывать.SG
‘…мучительная для него роль, потому что рассказывать он не умеет’

В русском языке при модальных глаголах также распространено опущение, причём сразу двух типов, ср.:

(17) *Умеешь ты обидеть!*
(18) *Он умеет только убивать.*

В первом случае конструкция с совершенным видом глагола акцентирует внимание на частоте события. Во втором более нейтральная конструкция с несовершенным видом подчёркивает именно умение говорящего. Именно поэтому в конструкции с совершенным видом, в отличие от конструкции с несовершенным, затруднено употребление обстоятельств типа *хорошо*:

(19) *Он хорошо умеет убивать / обижать людей.*
(20) *?Он хорошо умеет обидеть.*

С этим же различием связан тот факт, что при допустимости (21) странным выглядит (22): при совершенном виде глагола подразумевается, что событие частотно и, следовательно, уже имело место:

(21) *Вообще-то он умеет обижать людей, но никогда этого не делал / не делает.*
(22) *?Вообще-то он умеет обидеть, но никогда этого не делает.*

Впрочем, в русском языке случаи опущения статистически незначимы даже в пределах этих конструкций (во всяком случае, среди первых 200 примеров конструкции «*уметь* + инфинитив» встретилось всего два примера с глаголами несовершенного вида и один — с глаголом совершенного, в частности, (23) и (24)):

(23) А ты не обижайся. Я не умею обижать. Сердце у него билось мощно и ровно, постепенно успокаиваясь, и она слушала разгоревшимся ухом, лежа у него на руке. [Г. Я. Бакланов. В месте светлом, в месте злачном, в месте покойном (1995)]

(24) Она умеет выслушать, понять, дать добрый и дельный совет, поддержать. [Всегда в поиске (2001) // «Наука в Сибири» (Новосибирск), 2001.03.07]

Однако если рассмотреть только глаголы каузации эмоций (например, только глагол *обидеть*), то при поиске в системе Google из 14 примеров с этим глаголом только в одном выступает объект.

Именно контекст модального глагола (или сходные с ним контексты, где ситуация не является реальной) — это наиболее частотный фактор опущения объекта при глаголе *убивать* (23 из 200 примеров):

(25) Чтобы не боялись крови и не знали жалости. Чтобы умели убивать одним ударом. Чтобы учились ценить только собственную жизнь. [Борис Васильев. Вещий Олег (1996)]

Такая связь между модальной семантикой и опущением объекта вполне объяснима. Если говорится не о некоторой имеющей место ситуации P, а о ситуации, которая *может* возникнуть, ясно, что конкретизировать объект в таком случае говорящему не нужно, а иногда и невозможно. Изначально неясно, с каким именно участником будет иметь место ситуация. Отметим также, что, согласно [Hopper, Thompson 1980], конструкции со значением реальной ситуации и типологически обладают большей семантической транзитивностью, чем обозначающие возможную или ирреальную ситуацию.

3.1.3. Опущение нереферентного объекта в неактуальной ситуации

Другим фактором опущения объекта является неактуальный контекст: ситуация почти полностью становится свойством, переставая быть актуальной ситуацией. Сравним, например, предложения *Он стал преступником. Он грабит и убивает* и *В Турции за это убивают*. В первом предложении ситуация характеризует субъекта, но в то же время речь в предложении идёт о конкретных его действиях. Во втором предложении говорящий скорее рассказывает не о реально происходящих событиях, а о свойстве Турции, при этом реальных случаев убийства могло быть не так уж много.

В частности, показателем неактуальной ситуации является использование неопределённо-личной конструкции с глаголом в третьем лице множественного числа. В девяти случаев из выборки на глагол *убивать* мы видим именно эту конструкцию:

(26) «Что вы! — закричали турки. — В Турции жаловаться нельзя, в Турции за это убивают». [Фазиль Искандер. Дедушка (1966)]

Объяснение этих случаев сходно с предложенным для модального контекста. Если, например, свойством некоторого места является возникновение в нём ситуации P (‘убивать’), понятно, что в реальности ситуация P может возникать в разное время с разными объектами.

В то же время интересно, что в арабском языке сходных конструкций не обнаруживается вообще. Русским неопределённо-личным конструкциям по смыслу соответствуют пассивные. Следовательно, один из факторов опущения для арабского не является релевантным.

3.2. Типы опущения, нехарактерные для арабского языка: референтное опущение и опущение объекта-участника коммуникации

Одним из самых распространённых типов опущения в русском языке является опущение объекта-участника коммуникации. В нашей выборке с глаголом *убить* оно встретилось 14 раз. При этом надо сказать, что опущение референтных объектов, не являющихся участниками коммуникации, в нашей выборке не встретилось.

(27) — Можешь!! — взревел трактирщик. — Убью!.. — Убивай... — Застрелю! Сегодня же... [Василь Быков. Камень (2002)]

В примере (27) трактирщик имеет в виду ‘убью тебя’, а в ответной реплике, естественно, подразумевается ‘убивай меня’, то есть опущены местоимения, относящиеся, соответственно, к адресату и говорящему.

Если предыдущие типы опущения (3.1.2, 3.1.3) объяснялись тем, что соответствующий объект не был релевантен, определён и мог различаться для различных моментов, когда возникает ситуация, то в данном случае дело обстоит наоборот. Участники коммуникации строго определены, и зачастую обозначающие их местоимения можно опустить без ущерба для смысла.

Тем важнее тот факт, что в арабском языке рассматриваемый тип опущения встречается редко. Примеров типа *Удивляет тот факт, что никто даже не извинился* в арабском языке почти не встречается (об исключениях мы скажем ниже).

Это неудивительно, учитывая особенности маркирования прямого объекта в этом языке. Прямообъектные местоимения безударны, составляют единое фонетическое слово с глаголом и на письме пишутся с ним слитно. По своим свойствам они близки к глагольным аффиксам — неудивительно,

что они с трудом подвергаются опущению. Впрочем, и это правило знает исключения.

В частности, иногда встречается опущение объекта-участника коммуниации. Так, глагол *'amkana* 'быть возможным' в качестве прямого дополнения присоединяет экспериенцера (ср. *yumkinuni: 'an 'adkhul-a* 'возможно-мне войти'). Однако это дополнение может опускаться, хотя практически всегда определено, известно и в некоторых случаях совпадает с говорящим или слушающим:

(28) *hal y-umkin-u 'an n-atawaqqa'-a dalika*
Q 3SG.M-быть.возможным-SG чтобы 1PL-ожидать-CONJ тот. M.SG
'Можно ли этого ждать?'

(29) *hal y-umkin-u 'an n-uḥaddid-a raqm-an li-aS-Sa:dira:t-i*
Q 3SG.M-быть.возможным-SG чтобы 1PL-определить-CONJ номер-ACC для-DEF-выпуск.PL-GEN
'Можем ли мы определить число выпусков'

Число конструкций с выраженным Экспериенцером даже меньше. В основном, они встречаются в газете Al-Hayat (13 примеров) и Al-Ahram (6):

(30) *hal y-umkin-u-na: 'an n-ajid-a waSaf-an daqi:q-an?*
Q 3SG.M-быть.возможным-SG-1PL чтобы 1PL-найти-CONJ описание.SG-ACC точный-ACC.SG.M
'Можем ли мы найти точное описание?'

Правда, в данном случае конструкция с невыраженным и с выраженным объектом, как кажется, различаются семантически. При выраженном объекте часто (хотя и не всегда) речь идёт о том, **разрешено ли** действие (ср. русское 'Можно (нам, мне)').

При невыраженном объекте речь может идти о **возможности** действия, связанной с внешними обстоятельствами или внутренними свойствами Экспериенцера ('мы'), но не обязательно с разрешением (ср., впрочем, (30), где при выраженном объекте выражается, тем не менее, возможность).

Также может не выражаться объект при глаголах звука (*daqqa* 'стучать', *daqqa al-ba:b-a* 'стучать в дверь'). Однако эти лексемы, как показано в [Летучий 2006], вообще во многих языках имеют вариативную модель управления (начиная от собственно лабильности и заканчивая вариативностью выражения объекта).

Наконец, глаголы движения также могут не выражать прямое дополнение. Так, глаголы *waSala* 'прибывать', *dakhala* 'входить', *kharaĝa* 'выходить' в арабском переходные и управляет названием конечной точки в винительном падеже. Это дополнение часто не выражается:

(32) *'indama waSal-a Muhammad-un…*
когда прибыть-PST.3SG Мухаммад-NOM.SG
'Когда приехал Мухаммад…'

Свойства глаголов движения и звука заставляют нас ввести ещё одно противопоставление на множестве глаголов. Как оказывается, существенна семантическая роль объекта того или иного глагола — ниже мы подробнее рассмотрим этот вопрос.

4. Семантическая роль опускаемого объекта

Как показано в [Белова 1985], [Летучий 2006], особенностью арабского языка является то, что в позиции прямого дополнения может выступать более широкий класс объектов, чем в русском и во многих европейских языках. Иными словами, семантическая роль объекта почти не ограничена. Ср., например:

(34) *y-usa:w-i si'r-u ad-du:la:r-i si'r-a yu:ru*
3SG.M-равняться-SG цена-NOM.SG DEF-доллар-GEN цена-ACC.SG евро.NOM.SG
'Цена доллара равняется цене евро'.

Стативный глагол *sa:wa:* 'равняться' здесь является переходным и имеет прямой объект со значением Эталона сравнения — семантическая роль, которая в русском языке обычно не выражается с помощью переходной модели. Согласно [Hopper, Thompson 1980], наиболее прототипическим случаем переходного глагола является глагол с прямым объектом-Пациенсом (о свойствах Пациенса см. [Dowty 1991]).

С тем же случаем мы имеем дело в случае глаголов звука и движения. Прямым дополнением у них является Место — семантическая роль, для которой в целом так выражаться нехарактерно. Как правило, участники с ролью Места в различных языках не являются обязательными для выражения. В арабском имеет место противоречие между семантикой роли и её способом выражении: се-

мантически периферийная роль кодируется дополнением с высоким статусом. Для опущения играет роль, прежде всего, семантика: несмотря на свой высокий статус, дополнение легко может опускаться (возможно, то же имеет место при модальном глаголе *'amkana*).

Отметим, что в русском языке переходные глаголы движения, например, *достигнуть*, в меньшей степени допускают опущение дополнения. В то же время многим арабским переходным глаголам в русском языке соответствуют непереходные (например, араб. *kharaĝa* 'выходить' переходный, а русский *выходить* непереходный), при которых непрямое дополнение может опускаться.

5. Явление, связанное с опущением объекта: вариативность управления

В данной части мы кратко рассмотрим явление, близкое к опущению объекта — вариативность управления объектом. Хотя явление это другой природы, нежели опущение, их объединяет то, что в одном из употреблений глагол имеет выраженный прямой объект, а в другом объект не выражен или отсутствует вообще).

Приведём вначале пример из русского:

(36) a. *Он пожертвовал двадцать тысяч на ремонт церкви.*
b. *Ради неё он пожертвовал своей комсомольской карьерой.*

В значении 'давать (деньги и т. д.) на какую-либо цель' глагол *жертвовать* является переходным и управляет прямым дополнением, а также непрямым с предлогом *на*. При смене значения на 'соглашаться на потерю чего-л. взамен на что-л. другое' модель управления полностью меняется: жертвуемый объект теперь маркируется не винительным, а творительным падежом, а объект со значением Цели присоединяется с помощью предлога *ради*, а не *на*.

В арабском языке вариативность зачастую не нагружена семантически, ср., например, глагол *kašafa* 'открывать, обнаруживать', способный присоединять Пациенс в качестве прямого дополнения или с помощью предлога *'an* 'из, об'.

В данной статье не будем в полной мере рассматривать вариативность управления. Для нас важно, прежде всего, следующее. Корпусной материал позволяет заметить, что выбор того или иного варианта связан с грамматическими характеристиками глагола. В зависимости от породы, которая выбрана для данного контекста, управление может меняться: непроизводный глагол первой породы *kašafa* в форме 3 лица единственного числа женского рода *kašafat*[5] в 736 случаях управляет предлогом *'an* (*kašafat 'an sirr-i-ha:* 'она открыла её тайну, букв. о её тайне') и в 171 — прямым объектом (*kašafat-hu* 'она его открыла'). Напротив, его дериват восьмой породы *iktašafa*, практически не отличающийся по значению, в той же форме всего в 11 случаях присоединяет *'an* (*iktašafat 'an liqa:'-i:* 'она узнала о моей встрече (букв. открыла о моей встрече)') и в 111 — прямой объект (*iktašafat-ha:* 'она о ней узнала').

Данный результат, достигнутый средствами корпуса, довольно любопытен. Форма восьмой породы имеет чаще всего медиальное, рефлексивное или декаузативное значение (ср. *ghasala* 'мыть' — *ightasala* 'мыться'), то есть используется для понижения переходности глагола. В нашем же случае оказывается, что данная форма чаще является синтаксически переходной, чем исходная[6].

При этом сформулировать эту тенденцию, не используя статистические методы, нельзя. Описывая теоретически возможные модели, мы должны были бы сказать, что исходный глагол *bada'a* и производный от него глагол восьмой породы *iktašafa* имеют и переходное, и предложное управление.

Для русского языка такого рода случаи не слишком актуальны, поскольку реже наблюдается варьирование управления. Однако сходные случаи, когда общие синтаксические и семантические свойства глагольной модели не соблюдаются в ряде конкретных случаев, характерны и для русского. Ср., например, глаголы *плакать* и *плакаться*. Первый из них имеет одну семантическую валентность (Пациенс / Экспериенцер), у второго число валентностей и актантов возрастает до трёх (Пациенс / Экспериенцер, Содержание — *на что* и Адресат — *кому*). Это не согласуется с общими свойствами показателя *-ся*, который, как правило, понижает синтаксическую переходность и количество актантов глагола.

6. Текстовые различия

В ряде случаев материал корпуса позволяет выявить различия между местными разновидностями арабского языка (несмотря на то, что письменные тексты во всех арабских странах пишутся, в основном, на литературном языке). Так, глаголы «третьей породы» (с долгим гласным после первого согласного), как правило, обозначают симметричную ситуацию (например, *sa:wa:* 'равняться, уравнивать', *qa:raba* 'сближаться, приближать') и т. д.

---

[5] Данная форма выбрана потому, что, в отличие от формы мужского рода 3 лица единственного числа *kašafa*, не совпадает на письме (без использования огласовок) с формами отглагольных существительных.

[6] В настоящее время мы не можем сказать, насколько распространён данный случай.

У данных глаголов зачастую вариативное управление: так, глагол *sa:wa:* может управлять предлогом *bayna* ‘между’ (букв. ‘уравнивать между чем-л.’), а может — прямым объектом и объектом с предлогом *bi* ‘с’ (‘уравнивать что-л. с чем-л.’). В газете Al-Hayat, издающейся в Ливане и Саудовской Аравии, из 335 вхождений данного глагола в 70 встречается предлог *bayna*. Гораздо ниже его доля в египетской газете Al-Ahra m: всего 14 вхождений из 244 примеров с этим глаголом. Для газеты At-Tajdi:d доля тоже низка (4 из 48).

Различия наблюдаются и для глагола *qa:raba*. Для него, помимо того же предлога *bayna*, существует вариат с прямым объектом и объектом с предлогом *min* ‘из’ (‘приближать что-л. к чему-л.’, букв. ‘от чего-л.’). В кувейтской газете Al-Watan из 259 вхождений конструкция с *bayna* не встречается ни одного раза, конструкция с *min* очень частотна (56 примеров). Из 74 употреблений глагола в египетской газете Al-Ahram также ни разу не встречается вариант с *bayna*, но и конструкция с *min* наблюдается только в четырёх примерах. Вообще модель с *bayna* встречается, в основном, в сирийской газете Thawra.

Данные различия весьма симптоматичны. Считается, что переходность арабского глагола в значительной мере предсказывается его способом образования (в частности, в [Гранде 1998] напрямую указывается, что практически все глаголы третьей породы переходны). В действительности в случаях, когда семантическая транзитивность глагола невысока, и внутри литературного языка, и между диалектами наблюдается варьирование. Более того, в нашем случае мы видим различия даже не между диалектами, а между разновидностями литературного языка.

Заключение

Итак, в арабском языке наблюдается явление опущения прямого объекта. Однако набор его типов и контекстов отличается от русского.

Так, в отличие от русского, практически не встречается референтное опущение. Исключением являются случаи, когда опускается объект-участник коммуникации — в этом случае, опущение, в основном, охватывает модальные глаголы — и случаи опущения объекта при глаголах движения, звука и других глаголов, объект которых не является одушевлённым и не обладает свойствами прототипического пациенса по [Dowty 1991].

Наши данные показывают ценность корпусного подхода к изучению синтаксических процессов. Такой подход позволяет выявить тенденции, не отражённые в словарях (например, частотные характеристики опущений).

Интересным типологическим результатом является вывод, что способ выражения объекта не полностью обусловливает возможность или невозможность его опущения. Напомним, что местоименный прямой объект в арабском языке выражается с помощью клитического местоимения. Поскольку клитики весьма тесно связаны с глагольной формой, могло бы оказаться, что опускаться они не могут, однако, судя по примерам (28) и (29), где подразумевается объект-Говорящий (или группа, куда входит Говорящий), это не так.

Различия между русским и арабским языками показывают, что опущения в этих языках используется системой по-разному. Хотя, как было показано выше, все рассматриваемые типы опущения объяснимы семантически и / или прагматически, реально в двух языках каждому из типов отводится разное место. В русском языке опущение, прежде всего, связано либо с модальным или неактуальным контекстом (примеры (23)–(26)), либо является средством избежать выражения объекта-участника коммуникации (27) (поскольку такие объекты обычно определены, тематичны, и их выражение не несёт новой информации). Напротив, в арабском не выражаются, как правило, нереферентные объекты (ср. примеры (9), (10)) — объекты-участники коммуникации могут опускаться только у закрытого класса глаголов (см. (28), (29)), да и в целом примеров, где в качестве объекта очевидным образом нужно было бы восстанавливать участника коммуникации, довольно мало. Влияние модальных операторов на опущение в арабском языке, по всей вероятности, существует, однако проявляется слабее, чем в русском.

Наконец, арабский материал интересен тем, что позиция прямого объекта может быть занята актантами с самыми разными семантическими ролями (в русском этот набор уже). Однако оказывается, что для опущения зачастую существенна именно семантическая роль, а не синтаксический статус. Прямые объекты с непрототипическими для этой позиции ролями (Конечной точкой движения, Местом при глаголах звука) легко подвергаются опущению (так же, как в русском языке — их корреляты, непрямые объекты и обстоятельственные именные группы при глаголах типа *прибыть (в город)* или *стучать (в дверь / в барабан)*.

Литература

1. *Баранов Х. К. Арабско-русский словарь*. М.: Русский язык, 1996.
2. *Белова А. Г. Синтаксис письменных текстов арабского языка*. М.: Наука, 1985.
3. *Гранде Б. М. Курс арабской грамматики в сравнительно-историческом освещении*. М.: Восточная литература, 1998.
4. *Летучий А. Б.* Типология лабильных глаголов: семантические и морфосинтаксические аспекты. Канд. дисс. М.: РГГУ, 2006.
5. *Недялков В. П.* Некоторые вероятностные универсалии в глагольном словообразовании // Вардуль И. Ф. (ред.). Языковые универсалии и лингвистическая типология. М.: Наука, 1969. С. 106–114.
6. *Dowty D.* Thematic proto-roles and argument selection. Language, 67.3. 1991. P. 547–619.
7. *Haspelmath, M.* More on the typology of inchoative/causative verb alternations. In Comrie B. and Polinsky M. (eds). *Causatives and Transitivity*. Amsterdam/Philadelphia: Benjamins, 1993. P. 87–120.
8. *Hopper P. and Thompson S.* Transitivity in Grammar and Discourse. *Language*, 56.2. 1980. P. 251–299.
9. *Kazenin K. I.* On the Lexical Distribution of Agent-preserving and Object-preserving Transitivity Alternations. *Nordic Journal of Linguistics* 17. 1994. P. 141–154.
10. *McShane M. A Theory of Ellipsis*. Oxford University Press, 2005.

Пунктуационная структура художественных произведений и её роль в синтезе выразительной речи по тексту

Punctuation structure of works of art and its role in synthesis of expressive speech under the text

Лобанов Б. М. (lobanov@newman.bas-net.by),

Объединенный институт проблем информатики НАН Беларуси, Минск, Беларусь

В работе приводятся статистические сведения о частоте употребления различных знаков препинания, а также сведения о частоте употребления простых и сложных предложений в художественных произведениях различного жанра. Полученные статистические данные позволили выявить наиболее частотные знаки препинания и особенности их употребления, определить частотный и количественный состав простых и сложных предложений, сформулировать особенности использования информации о знаках препинания для просодического оформления синтезированной речи по тексту.

Введение

К настоящему времени системы синтеза речи достигли определённого совершенства и уже используются в ряде практических приложений, в том числе, при создании аудиокниг. Однако комфортность восприятия синтезированной речи в реальных условиях, в особенности в этом конкретном её применении, остаётся ещё далекой от удовлетворительной. Мировая тенденция развития речевых технологий указывает на актуальность создания систем синтеза выразительной речи (**expressive** text-to-speech). Число исследований в области создания систем синтеза выразительной речи постоянно увеличивается. Исследования в области создания систем синтеза выразительной речи проводятся для ряда европейских языков: английского [1], немецкого [2], а также для синтеза речи на японском языке [3]. Некоторые аспекты создания системы синтеза выразительной речи на русском языке рассмотрены в монографии [4].

Понятие «выразительность речи» в лингвистике и педагогике сформировалось как междисциплинарное понятие одной из функций устной речи человека [5]. Одним из главных компонентов звуковой реализации выразительности устной речи является просодика: мелодика, ритмика и динамика речи. Эти элементы взаимодействуют, поддерживают друг друга и все вместе обуславливают её выразительность. Пунктуационная структура текста отображается знаками препинания, которые выполняют функции выделения смысловых отрезков текста (предложений, словосочетаний, слов, частей слова), указания на грамматические и логические отношения между словами, указания на коммуникативный тип предложения, его эмоциональную окраску, законченность, а также некоторые иные функции. Знаки препинания, синтаксически оформляют текст, а при синтезе речи по тексту помогают осуществить её просодическое оформление (интонация, смысловые паузы, логические ударения).

На рисунке 1 представлена общая структура синтезатора выразительной речи. Синтез устной речи по тексту осуществляется на основе лексико-грамматического анализа входного текста с учётом правил произношения звуков и интонирования, свойственных данному языку. На рисунке отображены дополнительные блоки, содержащие наборы правил для лингвистического и просодического процессоров, которые необходимы при реализации синтеза выразительной речи.

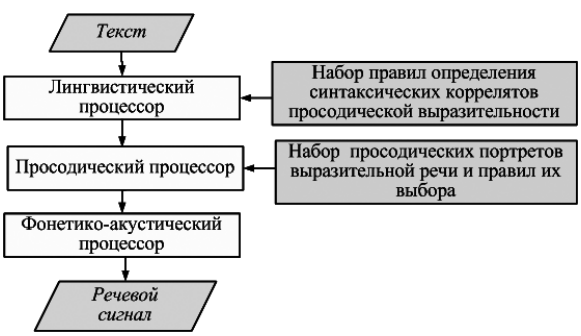


Рис. 1. Общая структура синтезатора выразительной речи

В данной работе мы коснёмся лишь одного из многих аспектов проблемы синтеза выразительной речи, а именно: выяснению, как воспользоваться информацией, заложенной в знаках препинания книжных текстов, для их выразительного «чтения» синтезатором речи? С этой целью мы проанализируем пунктуационную структуру 2-х произведений художественной прозы: Л. Толстой «Анна Каренина» и А. Толстой «Пётр Первый», а также полное собрание стихотворных произведений Пушкина и пьесу Гоголя «Ревизор». На примере этих произведений мы рассмотрим далее статистические характеристики и роль различных знаков препинания в выражении тех или иных интонационных свойств синтезируемой речи.

1. Статистика употребления знаков препинания в исследуемых текстах

В соответствии с особенностью исследуемой проблемы знаками препинания назовём явно выраженные в тексте специальные знаки (одиночные либо комбинированные), которые разделяют текст произведения на последовательность отрезков различной размерности: абзац, предложение, часть предложения, слово. К знакам препинания в порядке уменьшения размерности выделяемых ими отрезков текста отнесём следующие знаки:

- Знаки конца абзаца, отображаемые в тексте переводом строки + начальный отступ, либо двойным переводом строки. Первый тип абзаца (простой абзац) обозначим знаком **[||]**, второй тип абзаца (суперабзац) обозначим знаком **[|||]**. Простой абзац отделяет группу предложений внутри отдельной части (параграф, глава, часть), а суперабзац отделяет эти части произведения.
- Знаки конца предложения, отображаемые в тексте знаками: **[.]** — для предложений повествовательных, **[!]** — для восклицательных и **[?]** — для вопросительных предложений. Кроме этих основных знаков для выделения вводных предложений используются знаки **[( ]**, **[—]** — в начале и знаки **[ )]**, **[—]** — в конце вводного предложения. Знаками конца предложения могут быть также следующие комбинации из этих знаков: **[. —]**, **[…]**, **[…—]**, **[?!]**, **[? —]**, **[?…]**, **[! —]** **[!!!]**, **[!…]**. Кроме того, в качестве особых предложений следует определить заголовки и подзаголовки, не помечаемые зачастую ни одним из перечисленных выше знаков, выделение которых осуществляется по определённым правилам.
- Знаки препинания внутри сложных предложений, отображаемые знаками **[;]**, **[:]**, **[,]**, **[—]**, **[,—]**, выражают ту или иную степень «завершённости — незавершённости» выделяемой части предложения. Кроме того, к числу своеобразных выделителей частей сложного предложения следует отнести открывающие и закрывающие кавычки: **[«]**, **[»]** — знаки цитации или иронии.

В таблице 1. приведены обобщённые статистические характеристики пунктуационной структуры исследуемых текстах, полученные с использованием стандартных средств Microsoft Office.

Таблица 1. Статистические характеристики исследуемых текстов

		Кол. слов в тексте	Кол. всех предл. в тексте	Средн кол. предл. в абзаце	Средн. кол. слов в предл.	Средн. кол. знаков путкт. в предл.	Кол. поверств. предл. в тексте (%)	Кол. воскл. предл. в тексте (%)	Кол. вопр. предл. в тексте (%)
1	Л. Толстой «Анна Каренина»	280 336	23 223	3,1	12,1	2,3	83,3	6,7	10
2	А. Толстой «Пётр Первый»	229 192	28 574	5,6	8,0	1,6	91,1	4,0	4,9
3	А. Пушкин Стихи, поэмы.	208 320	26 333	3,7	7,9	1,2	78,5	12,1	9,4
4	Н. Гоголь Пьеса «Ревизор»	20 611	4365	3,9	4,7	1,0	74,6	15,5	9,9

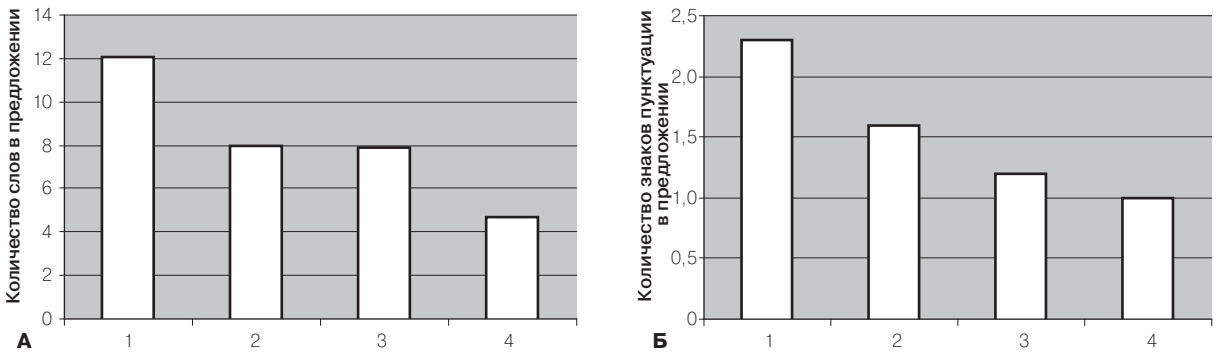


Рис. 2. Среднее количество слов (А) и знаков препинания (Б) в предложениях для текстов (1–4)

Как видно из данных таблицы 1 статистические характеристики встречаемости знаков препинания заметно различаются, как у текстов различного жанра (проза, стихи, пьеса), так и у текстов одного жанра (проза), но разных авторов. Наиболее значительные различия наблюдаются в значениях среднего количества слов (рис. 2А) и знаков препинания в предложениях (рис. 2Б) и в относительном количестве употреблений восклицательных и вопросительных предложений (см. рис. 3).

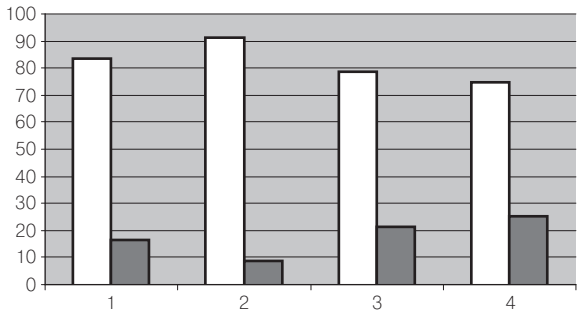


Рис. 3. Относительное количество (в %) повествовательных (ряд 1), восклицательных и вопросительных (ряд 2) предложений в текстах 1–4

Из рис. 3 видно также, что, хотя число повествовательных предложений существенно выше числа восклицательных и вопросительных, их количество всё же достаточно значительное и колеблется от 9 до 25%. Этот факт подчёркивает важность разработки правил синтеза восклицательной и вопросительной интонации (обычно недооцениваемая), в особенности, при создании систем синтеза речи по текстам стихотворного и драматического жанров.

2. Статистика употребления простых и сложных предложений

Статистические данные об употреблении в исследуемых текстах простых и сложных предложений получены с использованием разработанной в Лаборатории специальной программы, которая отбирает предложения, включающие хотя бы один знак препинания (условно сложные предложения) и предложения, в которых знаки препинания отсутствуют (простые предложения). В таблице 2 и на рисунке 4 приведены процентные соотношения сложных и простых повествовательных, восклицательных и вопросительных предложений в 4-х анализируемых текстах.

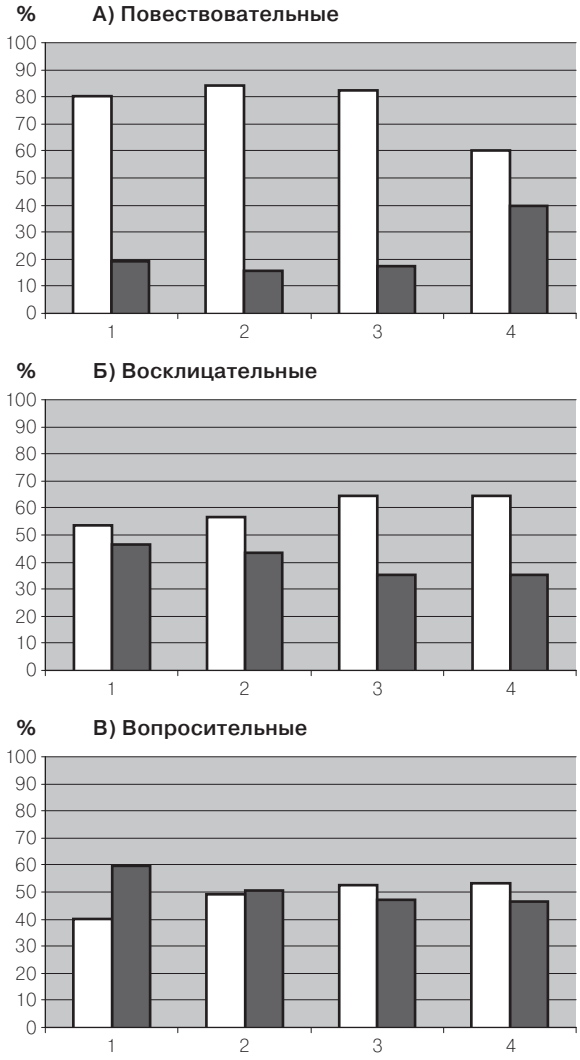


Рис. 4. Процентное содержание сложных (ряд 1) и простых (ряд 2) предложений в текстах 1–4

Как видно из рис. 4, повествовательные предложения в подавляющем большинстве представлены сложными предложениями (за исключением

Таблица 2. Статистические характеристики простых и сложных предложений

		Повеств. предл. [.]		**Восклиц. предл. [!]**		**Вопросит. предл.** [?]	
		сложн	простые	сложн	простые	сложн	простые
1.	«Анна Каренина» в %	80,6	19,4	53,6	46,4	40,1	59,9
2.	«Пётр Первый» в %	84,3	15,7	56,6	43,4	49,4	50,6
3.	Стихи Пушкина в %	82,6	17,4	64,6	35,4	52,8	47,2
4.	«Ревизор» Гоголя в %	60,1	39,9	64,7	35,3	53,5	46,5

4-го текста «Ревизор»). Превосходство относительного количества сложных восклицательных предложений над простыми сохраняется, хотя и в значительно меньшей степени, а для вопросительных — их количество для 1-го и 2-го текстов даже меньшее, чем число простых предложений.

Как известно, знаки препинания в большинстве случаев являются естественными маркерами просодического членения, используемыми при синтезе речи по тексту. Однако в простых предложениях они полностью отсутствуют. В то же время в естественной речи просодическая пауза, как правило, реализуется через каждые 2–4 слова даже при отсутствии знаков препинания. В связи с этим представляет интерес проанализировать количественную структуру простых предложений, представленных в выбранных текстах. В таблице 3 и на рисунке 5 приведены данные о среднем и максимальном количестве слов в простых повествовательных, восклицательных и вопросительных предложениях.

Как видно из рис. 5, в простых повествовательных предложениях прозаических произведений даже среднее число слов превышает четырёхсловный максимум, необходимый для реализации просодической паузы. Что же касается максимального количества слов в простых предложениях, зафиксированных в исследуемых произведениях, то оно значительно превосходит требуемые 4 слова и колеблется от 11 до 28 слов. Из рисунка 5 видна также существенная разница в распределениях, полученных для прозы (1, 2), стихов (3) и пьесы (4).

В таблице 4 приведены примеры простых и сложных предложений минимальной и максимальной длинны из романа Л. Н. Толстого «Анна Каренина», а в таблице 5 из стихов и поэм А. С. Пушкина.

3. Статистика употребления знаков препинания внутри и на концах предложений

В таблице 6 и на рисунке 6 приведены данные о среднем количестве «внутренних» знаков препинания **[,]**, **[—]**, **[, —]**, **[;]**, **[:]**, **[( ]**, приходящихся на одно сложное предложение, которое характеризуют частоту встречаемости знака. Иными словами, обратные значения чисел, приведенных в в таблице и на рисунках, показывают на какое количество предложений в исследуемых текстах приходится тот или иной «внутренний» знак препинания.

Как видно из приведенных данных, знак **[,]** наиболее частотный в каждом из текстов, чего нельзя сказать об остальных знаках препинания. Следующим по частоте встречаемости в прозаических произведениях является знак **[—]**, в то время как в стихотворениях Пушкина на второе место выходит знак **[;]**, а в пьесе Гоголя — знак **[( ]**.

Таблица 3. Среднее и максимальное количество слов в простых предложениях

		Повеств. предл. [.]		Восклиц. предл. [!]		Вопросит. предл. [?]	
		Средн.	Макс.	Средн.	Макс.	Средн.	Макс.
1.	Анна Каренина	5,5	28	2,7	14	3,7	18
2.	Пётр Первый	5,4	24	2,2	15	3,3	16
3.	Стихи Пушкина	3,2	21	2,8	22	4,3	21
4.	Ревизор “Гоголя”	2,5	16	2,8	12	3,6	11

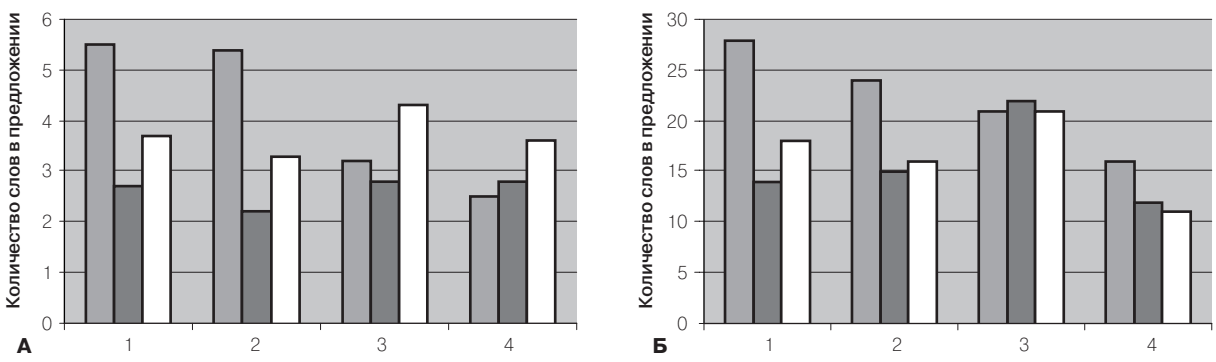


Рис. 5. Среднее (А) и максимальное (Б) количество слов в простых повествовательных (ряд 1), восклицательных (ряд 2) и вопросительных (ряд 3) предложениях

Таблица 4. Роман «Анна Каренина» Л. Н. Толстого

Тип предложения	Примеры
[.] простые	*Да. Анна. Никогда. Долли. Постой. Кити. Вронский. …* …… *И татарин с развевающимися фалдами над широким тазом побежал и через пять минут влетел с блюдом открытых на перламутровых раковинах устриц и с бутылкой между пальцами. Левин прислушивался к равномерно падающим с лип в саду каплям и смотрел на знакомый ему треугольник звезд и на проходящий в середине его Млечный Путь с его разветвлением.*
[.] сложные	*Да, доложи. Ну, хорошо. Ничего, батюшка. Напротив, очень……* …… *Ему казалось, что при нормальном развитии богатства в государстве все эти явления наступают, только когда на земледелие положен уже значительный труд, когда оно стало в правильные, по крайней мере в определенные условия; что богатство страны должно расти равномерно и в особенности так, чтобы другие отрасли богатства не опережали земледелия; что сообразно с известным состоянием земледелия должны быть соответствующие ему и пути сообщения, и что при нашем неправильном пользовании землей железные дороги, вызванные не экономическою, но политическою необходимостью, были преждевременны и, вместо содействия земледелию, которого ожидали от них, опередив земледелие и вызвав развитие промышленности и кредита, остановили его, и что потому, так же как одностороннее и преждевременное развитие органа в животном помешало бы его общему развитию, так для общего развития богатства в России кредит, пути сообщения, усиление фабричной деятельности, несомненно необходимые в Европе, где они своевременны, у нас только сделали вред, отстранив главный очередной вопрос устройства земледелия.*
[!] простые	*Да! Ааа! Да! Нехорошо! Аяяй! Долли! Ужасно! Уехал! Ага! Стива! Эге! Так!* …… …… *Какое вы доброе дело сделали вчера нашему жалкому соотечественнику!* *И как я смел соединять мысль о чем-нибудь не невинном с этим трогательным существом!*
[!] сложные	*Ах, ужасно! Левин, наконец! Долли, милая! Человек, хересу! Браво, Вронский! …* …… *Так же буду сердиться на Ивана кучера, так же буду спорить, буду некстати высказывать свои мысли, так же будет стена между святая святых моей души и другими, даже женой моей, так же буду обвинять ее за свой страх и раскаиваться в этом, так же буду не понимать разумом, зачем я молюсь, и буду молиться, — но жизнь моя теперь, вся моя жизнь, независимо от всего, что может случиться со мной, каждая минута ее — не только не бессмысленна, как была прежде, но имеет несомненный смысл добра, который я властен вложить в нее!*
[?] простые	*Как? Я? Что? Тюрбо? А? Отчего? Что? Клуб? Здоров? Где? Да? Кто? Дома?* …… …… *И почему же и всякий не может так же заслужить пред богом и быть взят живым на небо? Неужели эта трогательная радость его при ее приближении была причиной охлаждения Анны Павловны?*
[?] сложные	*Выпей, хочешь? Верно, хорошо? Ну, что? Что, Анна? …* …… *Разве не молодость было то чувство, которое он испытывал теперь, когда, выйдя с другой стороны опять на край леса, он увидел на ярком свете косых лучей солнца грациозную фигуру Вареньки, в желтом платье и с корзинкой, шедшей легким шагом мимо ствола старой березы, и когда это впечатление вида Вареньки слилось в одно с поразившим его своею красотой видом облитого косыми лучами желтеющего овсяного поля и за полем далекого старого леса, испещренного желтизною, тающего в синей дали?*

Таблица 5. Стихи, поэмы А. С. Пушкина

Тип предложения	Примеры
[.] простые	*Самозванец. Курбский. Поляк. Мнишек. Расходятся. Кавалер. Дама……* …… *В теченьи медленном река Вблизи плетень из тростника Волною сонной омывала И вкруг него едва журчала При легком шуме ветерка.* *Но в тишине души своей Она любви еще не знала И независимый досуг В отцовском замке меж подруг Одним забавам посвящала.*
[.] сложные	*Скучно, грустно. Делами, службой. Привычка, Ленский. …..* …… *Цветы, луга, ручей живой, Счастливый грот, прохладны тени, Приют любви, забав и лени, Где с Анжеликой молодой, С прелестной дщерью Галафрона, Любимой многими — порой Я знал утехи Купидона.*
[!] простые	*Ах! Ну! Ура! Парнаса! О! Изменница! Увы! Несчастный! Беги! Чу! ……..* …… *С каким тяжелым умиленьем Я наслаждаюсь дуновеньем В лицо мне веющей весны На лоне сельской тишины!* *Не дай мне бог сойтись на бале Иль при разъезде на крыльце С семинаристом в желтой шале Иль с академиком в чепце!*
[!] сложные	*Знай, Наталья! Но, лира! Ага, Мамон! Стой, путник! Молчи, молчи*! …… …… *А ты, повеса из повес, На шалости рожденный, Удалый хват, головорез, Приятель задушевный, Бутылки, рюмки разобьем За здравие Платова, В козачью шапку пунш нальем — И пить давайте снова!*
[?] простые	*Что? Краснеешь? Ужель? Куда? Доволен? Он? Кто? Деньги? Как? …….* …… *Когда в сияньи возгорится Светильник тусклый юных дней И мрачный путь мой озарится Улыбкой спутницы моей?* *Да как же ты не поспешил Тот час во след неблагодарной И хищникам и ей коварной Кинжала в сердце не вонзил?*
[?] сложные	*Ну, что? Я, государь? Что, какова? Что, маменька? Где вы, товарищи? ……* …… *Давно ли с трепетом народы Несли мне робко дань свободы, Знамена чести преклоня; Дымились громы вкруг меня, И слава в блеске над главою Неслась, прикрыв меня крылом?*

Таблица 6. Среднее количество знаков препинания внутри предложений

	[,]	[—]	[, —]	[;]	[:]	[( ]
«Анна Каренина» Л. Н. Толстого	**1,69**	**0,43**	**0,12**	**0,05**	**0,03**	**0,02**
«Пётр Первый» А. Н. Толстого	**1,01**	**0,37**	**0,1**	**0,07**	**0,08**	**0,04**
Стихи, поэмы А. С. Пушкина	**0,78**	**0,06**	**0,01**	**0,14**	**0,1**	**0,06**
Пьеса Н. В. Гоголя «Ревизор»	**0,64**	**0,04**	**0,05**	**0,03**	**0,06**	**0,14**

Рассмотрим далее статистику употребления знаков препинания на концах предложений: **[.], [||], [!], […], [?], [|| –], [ )], [. —], [? —], [… —], [! —], [!..].**

В таблице 7 и на рисунке 7 приведены данные о процентном соотношении предложений, маркированные различными знаками препинания для каждого из 4-х исследуемых текстов в порядке убывания средних значений по всем текстам. Как видно из приведенных данных наибольшим числом предложений со знаком **[.]** характеризуется прозаический текст 2 (Пётр I). Наибольшим количеством предложений со знаками абзаца **[||]**, **[!]**,**[?]** и **[…]** характеризуются стихотворный и драматический тексты 3, 4. Наибольшим количеством предложений со знаком **[|| —]** (индикатором диалоговой речи) характеризуется прозаический текст 1 (Анна Каренина). Другие особенности употребления знаков конца предложения можно увидеть при более детальном рассмотрении диаграмм (А–Г) на рисунке 7.

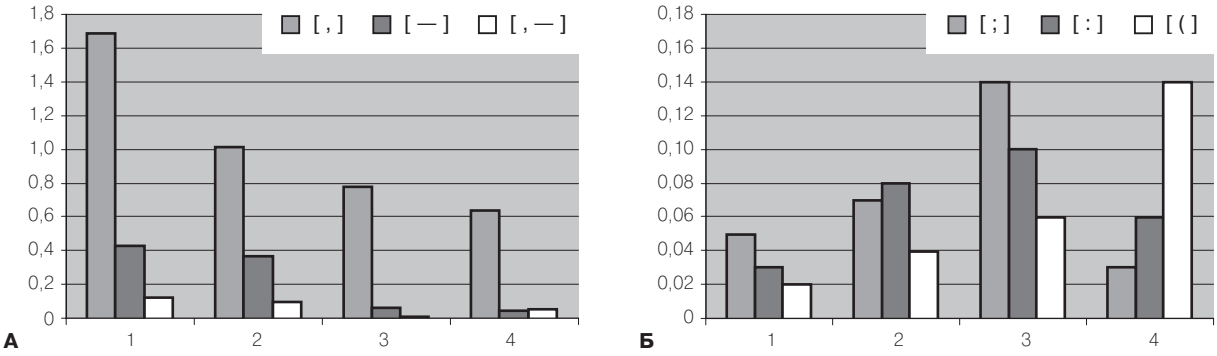


Рис. 6. Среднее количество знаков препинания [,], [—], [, —] — (А) и [;], [:], [( ] — (Б) внутри сложных предложений четырёх текстов

Таблица 7. Процентные соотношения предложений с различными знаками препинания

	[.]	[| |]	[!]	[...]	[?]	[| | —]	[ )]	[. —]	[? —]	[... —]	[! —]	[!..]
1	74	12	5,4	3,4	7,3	20	0,7	4,9	2,7	0,4	1,3	0
2	81	8	5,4	8,7	4,5	10	1,9	1,3	0,4	0,1	0,6	0
3	68	26	11,8	10	9,2	0,6	2,1	0,3	0,2	0,1	0,3	0
4	60	26	15,2	13	9,3	0	7,4	0,1	0,6	2,4	0,3	1,2
Средн.	70,7	18,0	9,5	8,8	7,6	7,5	3,0	1,6	1,0	0,7	0,6	0,3

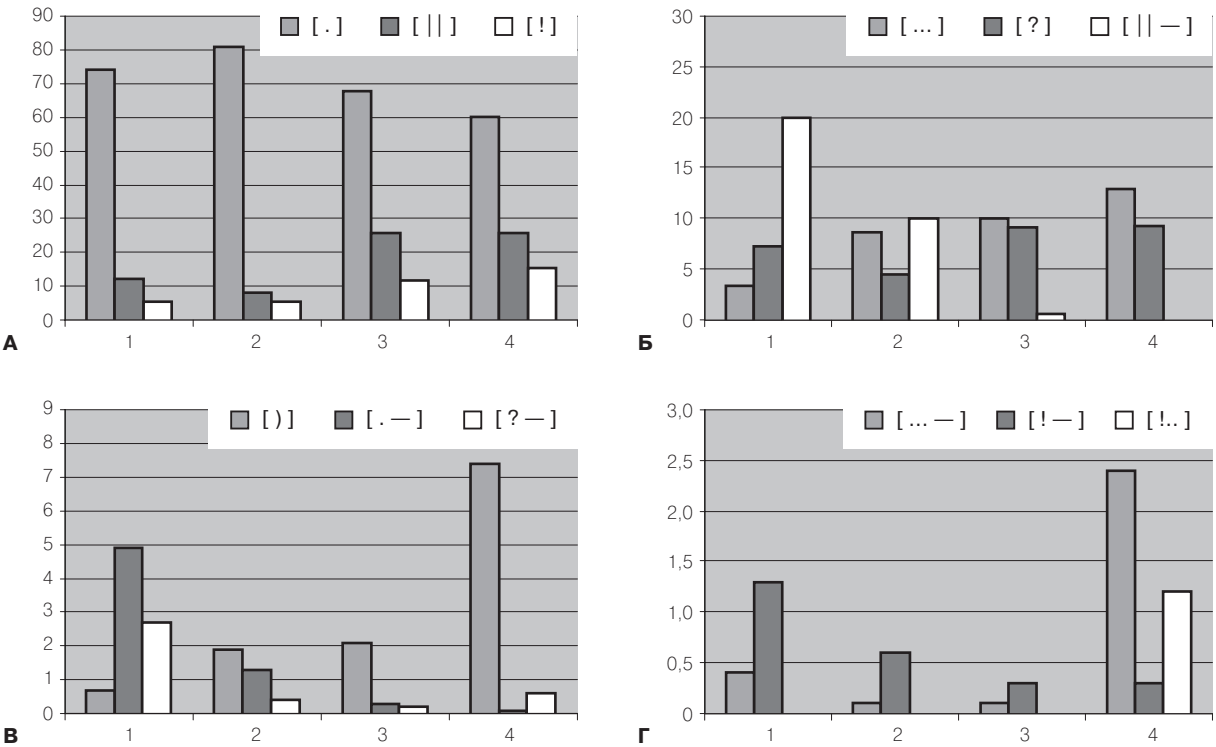


Рис. 7. Процентное соотношение предложений со знаками [.], [| |], [!] — (А), знаками [...], [?], [| | —] — (Б), знаками [ )], [. —], [? —] — (В) и знаками [.... —], [! —], [! ...] — (Г) для четырёх текстов

4. Использование информации о знаках препинания при синтезе выразительной речи по тексту

Как уже было отмечено, знаки препинания при синтезе речи по тексту потенциально могут способствовать осуществлению просодического оформления синтезированной речи, помогая выбору вариантов интонирования, паузации и логического ударения. Безусловно, этой формальной информации не достаточно в полной мере для синтеза выразительной речи, однако в отсутствии понимания текста при его машинном чтении, информация о пунктуационной структуре текста приобретает исключительное значение и должна быть использована в полной мере.

Прежде всего, каждый из знаков препинания можно считать индикатором членения текста на, так называемые [5], пунктуационные синтагмы, которые в синтезированной речи отделяются паузой той или иной длительности (от 0 до нескольких секунд). В порядке уменьшения относительной длительности межсинтагменной паузы знаки препинания можно условно разбить на следующие группы:

1. (2–5 сек) — суперабзац, отделяющий отдельные части произведения (параграф, глава и др.) — **[|||]**.
2. (1–2 сек) — абзацы, соответствующие разделам повествовательной — **[||]** и диалоговой речи **[|| —]**.
3. (0,5–1 сек) — повествовательные — **[.]**, **[. —]**, **[…]**, **[…. —]**, восклицательные — **[!]**, **[! …]**, **[!!!]** и вопросительные предложения — **[?]**, **[? —]**, **[?!]**.
4. (0,2–0,5 сек) — вводные предложения — **[( )]**, **[— / —]**, а также части предложения, отделяемые знаками — **[;]**, **[:]**, **[—]**.
5. (0–0,2 сек) — части предложения, отделяемые запятыми. Примером нулевой длительности паузы является запятая перед вводным словом: « *И, наконец, он сказал*».

Знаки препинания внутри и на концах предложений кроме длительности межсинтагменной паузы в значительной степени определяют также интонационный тип каждой из пунктуационных синтагм.

Знаки препинания на концах предложений: **[|||]**, **[||]**, **[.]**, **[. —]**, **[ )]**, **[…]**, **[…—]** — определяют тот или иной подтип интонации завершенности, знаки: **[!!!]**, **[!]**, **[! —]**, **[!..]** — тот или иной подтип восклицательной интонации, а знаки: **[?!]**, **[?]**, **[? —]** — вопросительной.

Знаки препинания внутри предложений: **[;]**, **[:]**, **[( ]**, **[,]**, **[—]**, **[, —]** — определяют тот или иной подтип интонации незавершенности. Особенно разнообразную роль в реализации интонации незавершенности играют знаки **[,]** и **[—]**.

Запятая может играть различную функциональную роль в предложении:

- отделение однородных членов предложения;
- отделение предикативных частей бессоюзного сложного или сложносочиненного предложения;
- отделение главного и придаточного предложений;
- при выделении обособленных членов предложения (причастий, деепричастий).

В зависимости от этой роли реализуется тот или иной подтип интонации незавершенности.

Тире в предложении также употребляется в разнообразной функциональной роли:

- отделение подлежащего и сказуемого в эллиптических предложениях;
- отделение однородных членов предложения перед обобщающим словом;
- для выражения неожиданности, резкой противоположности между членами предложения;
- для выделения вводных и вставных предложений;
- для выделения слов автора при прямой речи.

Заключение

В данной работе получены статистические сведения о частоте употребления различных знаков препинания, а также сведения о частоте употребления простых и сложных предложений в 2-х произведениях художественной прозы (Л. Толстой «Анна Каренина», А. Толстой «Пётр Первый»), в стихотворных произведениях А. Пушкина и в пьесе М. Гоголя «Ревизор». Полученные статистические данные позволили выявить наиболее частотные знаки препинания и особенности их употребления в художественных произведениях различных жанров, определить особенности структуры и количественный состав простых и сложных предложений. Сформулированы некоторые особенности использования информации о знаках препинания для просодического оформления синтезированной речи по тексту.

Литература

1. *Pitrelli J. F.* at al. The IBM expressive text-to-speech synthesis system for American English // Audio, Speech, and Language Processing, IEEE Transactions on V. 14, Issue 4, July 2006 Page(s):1099–1108
2. *Froehlich P., Hammer F.* Expressive Text-to-Speech: A user-centred approach to sound design in voice-enabled mobile applications // Telecom. Research Centre Vienna, Austria / http://userver.ftw.at/~froehlich/papers/JDS2004_ExpressiveTTS.pdf)
3. *Kawanami H.* Designing Speech Database with Prosodic Variety for Expressive TTS system // Nara Institute of Science and Technology, Takayama-cho, Japan / (http://gandalf.aksis.uib.no/lrec2002/pdf/337.pdf)
4. *Лобанов Б. М.* Компьютерный синтез и клонирование речи // Б. М. Лобанов, Л. И. Цирульник / Минск, Белорусская Наука.: 2008. — 342 с.
5. *Головин И. Б.* Основы культуры речи // Санкт-Петербург, Слово.: 1983

Квазисинонимы в лингвистических онтологиях

Near-synonyms in linguistic ontologies

Лукашевич Н. В. (louk@mail.cir.ru)

Научно-исследовательский вычислительный центр МГУ
им. М. В. Ломоносова; АНО Центр информационных исследований

Одной из важных проблем разработки лингвистических онтологий является вопрос представления значений квазисинонимов посредством набора дискретных понятий онтологии. В статье кратко рассматриваются подходы к представлению квазисинонимов в тезаурусе WordNet и онтологии МикроКосмос, а также подробно описываются принципы описания квазисинонимов в тезаурусе русского языка РуТез.

1. Введение

Для обработки текстов на естественном языке разрабатываются такие ресурсы как лингвистические онтологии, то есть онтологии, понятия которых в значительной мере связаны со значениями языковых единиц, терминов предметной области.

Одной из серьезных проблем разработки лингвистической онтологии является выработка принципов формирования единиц (понятий) онтологии. Понятия онтологии должны соответствовать понятиям предметной области [9], а, как известно, взаимоотношения между понятиями и языковыми значениями достаточно сложны.

Общие рекомендации по вводу понятий в онтологии заключаются в том, что понятие онтологии должно отчетливо отличаться от близких понятий в понятийной сети (родовых понятий, видовых понятий, понятий-сестер). Кроме того, нужно различать понятие и его названия: не стоит заводить отдельные понятия для синонимов [1, 6, 8]. Эти рекомендации не так просто выполнять, если онтология создается на основе реальных языковых значений [3].

Во-первых, непросто отличать понятия и его названия, работая с языковыми значениями. Во-вторых, серьезную проблему представляют совокупности близких по смыслу слов — квазисинонимов, значения которых различаются по нескольким характеристикам (понятийному содержанию, отношению говорящего, коллокациям и др.), видоизменяются в зависимости от контекста. Для многих таких совокупностей квазисинонимов чрезвычайно трудно установить однозначное соответствие на других языках, поскольку, чаще всего, на другом языке данной совокупности квазисинонимов соответствует другая совокупность квазисинонимов, которая характеризуется своей системой параметрических различий и соответственно своими особенностями.

Проблема представления значений квазисинонимов в лингвистической онтологии состоит в том, что не всегда ясно, как наилучшим образом представить совокупность близких значений квазисинонимов набором отдельных понятий. Лингвистическая онтология, которая хоть и учитывает существующие лексические значения, все же должна оставаться онтологией. По общим принципам организации онтологической иерархии основные элементы онтологии — понятия должны иметь четкие, независимые от контекста отличия от соседних понятий.

Создание различимых понятий в лингвистической онтологии важно еще и тем, что четкое понимание различий близких понятий позволяет аккуратно описать их отношения между собой и с соседними понятиями. Кроме того, различимые понятия онтологии легче описать средствами других языков.

В данной статье мы кратко рассмотрим подходы к представлению квазисинонимов в таких лингвистических онтологиях как WordNet [5] и MikroKosmos [6], а также подробно опишем принципы описания квазисинонимов в Тезаурусе русского языка РуТез [11], который мы также рассматриваем как лингвистическую онтологию.

2. Квазисинонимы в лингвистической онтологии WordNet

Наборы синонимов — синсеты — являются основными структурными элементами тезауруса

WordNet [5]. Авторы данного тезауруса считают два выражения синонимичными и относят их к одному и тому же синсету, если замена одного из них на другое в большинстве предложений не меняет значения истинности этого высказывания.

Этот основной принцип устройства WordNet приводит к тому, что не выполняется один из важнейших принципов разработки онтологий — это различение собственно понятия и способов его называния, то есть вводятся разные синсеты для разных способов наименования одной и той же сущности.

Имеется несколько типов смешений понятий и их названий в ресурсах типа WordNet.

Во-первых, смешение понятий и их названий проявляется в поддержке разных иерархий для разных частей речи. Действительно, с помощью, какой бы части речи в тексте не было бы упомянуто понятие ПРИВАТИЗАЦИЯ (*приватизировать, приватизационный, приватизация*) — это всегда ссылка на одно и то же понятие разными лексическими средствами, от изменения части речи не должны меняться отношения этого понятия с другими понятиями.

Первые разработчики ворднетов для других языков в рамках проекта EuroWordNet рассматривали возможность соединения всех дериватов в одной иерархии, поскольку такое разделение противоречит принципам разработки онтологических ресурсов. Однако, в конце концов, решение о соединении частей речи принято не было [2].

Вторым типом проявления смешения понятия и его названия является использование разных синсетов для описания старых и новых названий, названий понятия в разных диалектах языка, в разных текстовых жанрах и т. п.

Следствием принципа синонимичной подстановки является то, что WordNet имеет значительное количество синсетов, которые плохо отличимы друг от друга, что также нарушает онтологические принципы описания понятий. Так, например, имеется четыре различных синсета, обозначающие *сходство, подобие*, каждый следующий из которых является гипонимом для предыдущего и при этом является практически не отличимым от своего гиперонима:

sameness **—** *(the quality of being alike; "sameness of purpose kept them together")*

similarity *—* *(the quality of being similar)* — сходство

likeness, alikeness, similitude *— (similarity in appearance or character or nature between persons or things; ``man created God in his own likeness'')* — сходство по внешности, характеру или природе между людьми или объектами).

resemblance *— (similarity in appearance or external or superficial details)* — сходство во внешности или во внешних или поверхностных деталях.

3. Квазисинонимы в онтологии МикроКосмос

В лингвистической онтологии МикроКосмос [6] собственно онтология и лексикон разделены. Лексикон системы описывает значения слов и словосочетаний, устанавливая ссылки на понятия онтологии. Проблема квазисинонимов решается за счет объединения квазисинонимов к одному и тому же понятию онтологии, особенностей конкретных лексем описываются в словарных статьях словаря.

Авторы онтологии приводят пример, что все глаголы изменения в онтологии приписаны одному и тому же понятию CHANGE-EVENT [7]. Особенности слов описываются в словарной статье, например, для глагола *увеличить (increase*) указывается, что в семантической роли ТЕМА этого глагола должна выступать СКАЛЯРНАЯ_ВЕЛИЧИНА (например, цена или высота), и указывается, что значение этой величины меняется на большее.

Если мы обратимся к сайту ресурса, то мы увидим, что ситуация с реализацией изложенных принципов достаточно сложная. Так, понятию CHANGE_EVENT сопоставлен в лексиконе большой список слов, которые, по мнению авторов, онтологии соответствуют этому понятию, например: *acclimatization* (акклиматизация — приспособление к другому климату), *commerzialization* (коммерциализация), *contamination* (загрязнение), *damage* (повреждать), *deteriorate* (ухудшать), *improve* (улучшать) и многие другие — для этих слов не было заведено отдельных понятий.

В то же время среди нижестоящих по иерархии понятий можно увидеть следующие: ADJUST (адаптировать, приспособить), CORRECT-EVENT (исправление, коррекция), DIVIDE (делить), INTEGRATE (интегрировать), RESTRUCTURE (реструктуризация) и др. Непонятно, почему для одних значений слов были заведены отдельные понятия, а для других нет. Почему значение слова *acclimatization* не заслуживает отдельного понятия, хотя есть важное отношение к климату, биологическим процессам, а значение слова *adjust* такой концепт получило?

Помимо вопросов последовательности/непоследовательности описания имеются и явные последствия для процедур автоматической обработки текстов.

Так, сложной становится процедура установления, какие все-таки слова из большего списка словарных входов к понятию CHANGE-EVENT, могут рассматриваться как синонимы, каковы соотношения между этими словами. Кроме того, относительно небольшая величина онтологии приводит к тому, что при работе в конкретном приложении и конкретной предметной области многое придется доделывать и вводить дополнительные понятия даже для слов, которые уже учтены в онтологии.

Таким образом, на наш взгляд, в приведенных примерах MikroKosmos проблема квазисинонимов

решается путем чрезмерного переобобщения, что может привести к проблемам в реальных предметных областях. Необходимо выделить дополнительный уровень понятий, который поможет более четко разделить слова, не сваливая их в единый, большой мешок.

4. Понятия и значения в тезаурусе русского языка РуТез

Наиболее точно «жанр» тезауруса РуТез можно охарактеризовать как лингвистическая онтология для автоматической обработки текстов, то есть это онтология, большинство понятий которой вводится на основе значений реально существующих языковых выражений.

Значения языковых выражений, которые могут породить отдельное понятие в тезаурусе РуТез, относятся не только к общей лексике, но и могут являться терминами конкретных предметных областей, относящихся к сфере общественной жизни (экономика, право, международные отношения, политика), к сферам обслуживания населения (транспорт, банки и др.). Это связано с тем, что жизнь конкретных групп населения значительно связана с некоторыми профессиональными сферами деятельности, многие понятия из этих сфер легко перетекают в сферу общего языка, могут начать обсуждаться в общезначимых источниках информации (газетах, новостных сообщениях).

В качестве источников понятий в тезаурусе РуТез также активно используются словосочетания. Основным принципом введения такого рода понятий является необходимость фиксации некоторой дополнительной информации, которую невозможно описать в понятиях, соответствующих значениям слов — компонентов словосочетания.

В тезаурусе РуТез единицей является не множество синонимичных слов или терминов как в тезаурусе WordNet, а понятие — как единица мышления, обобщающая предметы и явления действительности [9]. Для ссылки на понятие в тексте могут использоваться несколько синонимичных текстовых выражений. Слова и словосочетания, значения которых могут быть представлены как ссылки на одни и те же понятия тезауруса, будем называть онтологическими синонимами.

Таким образом, онтологическими синонимами могут являться:

- слова, являющиеся разными частями речи (*стабилизация, стабилизироваться, стабилизационный),*
- языковые выражения, относящиеся к разным языковым стилям (*коммунальная квартира, коммуналка*),
- отдельные слова, устойчивые выражения, свободные словосочетания, значения которых соответствуют данному понятию (*аэропорт- воздушные ворота, газ — газообразное вещество)*.

Каждое понятие в тезаурусе имеет понятное, однозначное имя, что важно для ведения тезауруса, анализа результатов автоматической обработки текстов. В качестве названия может выступать однозначное словосочетание, однозначное слово, или пара синонимов, пересечение значений которых однозначно идентифицирует данное понятие.

В настоящее время тезаурус РуТез включает более 52 тысяч понятий, к которым приписано более 160 тысяч слов и словосочетаний. Тезаурус используется для таких видов автоматической обработки текстов как автоматическое концептуальное индексирование, автоматическое рубрицирование, аннотирование, кластеризация.

5. Принципы представления значений квазисинонимов в тезаурусе РуТез

Поскольку в настоящее время понятия тезауруса РуТез не имеют внутренней структуры в виде фреймовых элементов или атрибутов, то отличительные свойства понятий могут проявляться в наборе отношений с другими понятиями или в особенностях ассоциированных с понятием онтологических синонимов.

Для описания набора близких по смыслу значений квазисинонимов посредством набора различимых понятий лингвистической онтологии в РуТез применяется следующая процедура, которую мы рассмотрим на примере синсетов из WordNet, отражающих значение сходства (см. п. 2)

На первом шаге необходимо выделить наиболее существенные для тезаурусного описания признаки квазисинонимов, то есть такие признаки, в зависимости от которых требуется установление разных отношений с другими понятиями тезауруса.

В совокупности английских слов со значением сходства (similarity), таким признаком, например, является способность выражать сходство по внешним характеристикам. Значения некоторых квазисинонимов этой группы часто применяются именно к внешним характеристикам предметов, то есть в тезаурусе должно быть обозначено отношение к соответствующему понятию:

likeness, alikeness, similitude — *(similarity in appearance or character or nature between persons or things; "man created God in his own likeness")* — сходство по внешности, характеру или природе между людьми или объектами.

resemblance — *(similarity in appearance or external or superficial details)*.

Это означает, что в языке, жизни людей значимым является сходство по внешним характеристикам и нужно отразить этот факт соответствующим понятием.

На втором шаге необходимо подыскать подходящее название такому понятию

В случае квазисинонимов к слову *similarity,* таким названием понятия может служить словосочетание *Similarity in appearance* (34700 страниц в поисковой системе Google). Понятие вводится в тезаурус с таким названием.

На третьем шаге необходимо найти разные способы выражения этого же понятия в виде словосочетаний и отдельных слов, например, *resemblance in appearance, similarity of appearance, external resemblance* и др. Все эти варианты добавляются в качестве текстовых вариантов к понятию (рис.1).

На четвертом шаге для отражения значений слов, которые часто выражают именно это понятие, но могут использоваться и для выражения сходства вообще, например, *resemblance*, такое слово указывается как текстовый вход к понятию SIMILARITY IN APPEARANCE и как текстовый вход к более общему понятию SIMILARITY

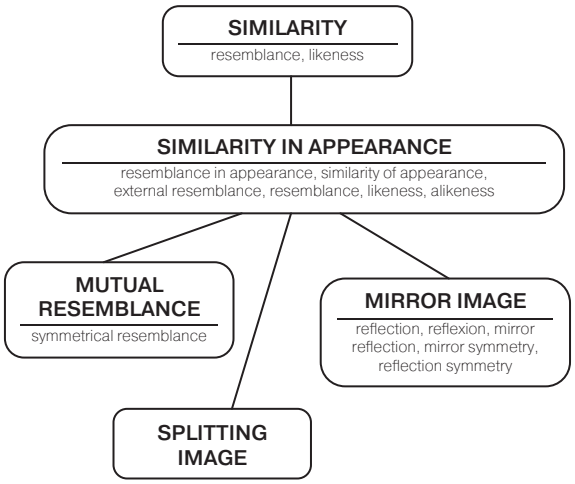


Рис. 1. Фрагмент совокупности отличимых понятий, отражающих значения квазисинонимов слова similarity. Заглавными буквами указаны имена понятий, прописными — соответствующие текстовые входы понятия. Текстовые входы, совпадающие с именем понятия, не указаны.

В случае, если независимых от контекста характеристик для различения значений квазисинонимов, найти не удается, то необходимо представить их в виде одного понятия. Для большей ясности имя такого понятия может быть составлено как пара соединенных в этом понятии синонимов.

В качестве примера анализа значений квазисинонимов на русском языке возьмем синонимические ряды, представленные в синонимическом словаре НОСС [12]. Этот словарь интересен тем, что его словарная статья содержит подробный перечень сходных черт и различий синонимов. На основе такой словарной статьи разбора удобно показать, какие различия приводят к представлению синонимического ряда словаря в виде онтологических синонимов одного и того же понятия, а для значений каких слов, представленных в данном словаре, как синонимы, введены несколько понятий, и, таким образом, в рамках тезауруса РуТез они синонимами не являются.

В качестве первого примера рассмотрим пару синонимов *памятник, монумент*.

В словаре НОСС [12, стр. 257] указываются пять различий употребления этих слов (по величине, форме, увековечиваемому объекту и др.). Анализ примеров употребления этих синонимов показывает, что указанные различия выполняются лишь по умолчанию, имеется достаточное число примеров употребления обоих синонимов в связи со всеми возможными типами увековечиваемых сущностей. Так, авторы словаря утверждают, что «в память о конкретном человеке обычно ставится памятник, о группе людей — и памятник, и монумент, о событии — монумент; идеи воплощаются в монументах».

Между тем, в память о конкретном человеке может быть установлен монумент:

Монумент выдающемуся исследователю севера Западной Сибири, лесоводу, этнографу Александру Дунину-Горкавичу торжественно откроется в Ханты-Мансийске. (http://ural.rian.ru/culture/20070614/81566803.html).

В память события может быть установлен памятник:

На Пролетарской площади вновь оборудован сквер, в котором установлен памятник Победы (http://www.megatula.ru/site/tulskii_krai/raionnye_centry/67/)

Памятник может быть поставлен идее:

Он сказал, что это не первая акция вандалов в отношении памятника русско-армянской дружбы (http://www.patriarchia.ru/db/text/56928.html)

Кроме того, авторы словаря указывают, что различия «нейтрализуются при повторной, сокращенной номинации того же сооружения». Таким образом, у слов *монумент* и *памятник* не нашлось ни одного четкого различающего свойства, которые привели бы к установлению разных отношений с другими понятиями тезауруса, поэтому эти два слова должны рассматриваться как онтологические синонимы.

В качестве второй пары синонимов, которую мы проанализируем с помощью словаря НОСС [12], рассмотрим пару слов *водитель, шофер*.

При рассмотрении этих слов авторы словаря указывают следующее различие: «шофер управляет только автомобилем или автобусом, водитель и другими транспортными средствами [12, стр.53]». Из этого замечания понятно, что *шофер* и *водитель* не могут быть онтологическими синонимами, поскольку водитель должен иметь отношения с понятиями, соответствующими словам *вагоновожатый, судоводитель*, а *шофер* — нет. Это означает, что для отражения значений этих слов необходим ввод, по крайней мере, двух понятий с названиями *ВОДИТЕЛЬ ТРАНСПОРТНОГО СРЕДСТВА* и *ВОДИТЕЛЬ АВТОМОБИЛЯ*. Видовыми понятиями для понятия *ВОДИТЕЛЬ ТРАНСПОРТНОГО СРЕДСТВА* будут такие понятия как *ВАГОНОВОЖАТЫЙ*, *СУДОВОДИТЕЛЬ*.

В то же время, носители языка ощущают эти слова как синонимы (см. также [10]). Чтобы отразить и это ощущение, и способность расширительного употребления, необходимо слово *водитель* представить как текстовый вход к двум понятиям *ВОДИТЕЛЬ ТРАНСПОРТНОГО СРЕДСТВА и ВОДИТЕЛЬ АВТОМОБИЛЯ*.

Сначала представляется, что слово *шофер* должно быть отнесено как текстовый вход к понятию *ВОДИТЕЛЬ АВТОМОБИЛЯ,* но можно заметить, что водители автомобилей могут быть любителями, и профессиональными работниками, а слово *шофер* все-таки относится к профессиональным водителям. Таким образом, онтологический анализ пары синонимов показал, что для адекватного отражения системы понятий, скрывающихся за близкими по смыслу словами *водитель* и *шофер*, нужно использовать три понятийные единицы: *ВОДИТЕЛЬ ТРАНСПОРТНОГО СРЕДСТВА, ВОДИТЕЛЬ АВТОМОБИЛЯ, ШОФЕР (ПРОФЕССИОНАЛЬНЫЙ ВОДИТЕЛЬ)* (см. рис. 2).

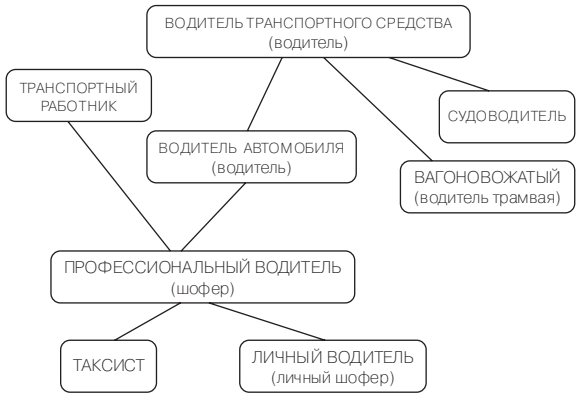


Рис. 2. Понятийная структура, соответствующая близким по значению словам водитель и шофер

Необходимость принятия решений о представлении значений близких по смыслу языковых выражений посредством совокупности понятий возникает и в конкретных предметных областях.

Так, ситуации кредитования соответствуют такие слова и словосочетания как: *кредитование, кредит, кредитная услуга, кредитное обслуживание, кредитная операция, выделение кредита, выдача кредита, выделение кредитных средств, предоставление кредита* и др. Имеется специфика употребления конкретных выражений из этого списка. Однако неправильным является введение дополнительных понятий онтологии для отражения именно специфики употребления. И в данном случае каждое вводимое понятие должно иметь четкий набор отличительных отношений. До тех пор, пока такие отличия не выделены, все такие выражения должны представляться как онтологические синонимы.

Заключение

Развивая тезаурус РуТез как лингвистическую онтологию, мы пытаемся следовать двум, вообще говоря, противоречивым критериям.

С одной стороны, мы формируем понятия тезауруса максимально близко к значениям языковых выражений, поскольку считаем, что чрезмерное обобщение, кластеризация значений ведет к искажению системы отношений, проблемам в приложениях автоматической обработки текстов.

С другой стороны, мы стараемся, чтобы понятие тезауруса было действительно понятием, то есть было отличимо от близких по смыслу понятий. Во многих случаях использованием реально существующих многословных выражений позволяет нам смягчить эти противоречивые требования. Введение понятия на базе значения многословного выражения не меняет суть лингвистической онтологии, но во многих случаях позволяет ввести более отчетливо отделимые понятия.

Литература

1. *Bouaud J., Bachimont B., Charlet J., Zweigenbaum P.* Methodological principles for structuring an “ontology”// Proceedings of IJCAI-95 Workshop “Basic Ontological Issues in Knowledge Sharing”, 1995.
2. *Climent S., Rodriguez H., Gonzalo J.* // Definitions of the links and subsets for nouns of the EuroWordNet project. — Deliverable D005, WP3.1, EuroWordNet, LE2-4003, 1996.
3. *Hirst G.* Ontology and the Lexicon // Staab S., Studer R. (eds.) Handbook on Ontologies in Information Systems. Berlin: Springer, 2003. P. 209–230.
4. *Magnini B., Speranza M.* Merging Global and Specialized Linguistic Ontologies // Proceedings of OntoLex 2002, 2002. С. 43–48.
5. *Miller, G., Beckwith, R., Fellbaum, C., Gross D., Miller K.* Five papers on WordNet // CSL Report 43, Cognitive Science Laboratory, Princeton University, 1990.
6. *Nirenburg S., Raskin V.* Ontological Semantics. MIT Press: 2004.
7. *Nirenburg S., McShane M., Beale S.* The Rationale for Building Resources Expressly for NLP. // Proceedings of the 4st International Conference on Language Resources and Evaluation (LREC). Lisbon, Portugal, 2004. P. 3–6.
8. *Noy N. F., McGuinness D.* Ontology Development 101: A Guide to Creating Your First Ontology // Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880, 2001.
9. *Smith B.* Beyond Concepts: Ontology as Reality Representation // Proceedings of International Conference on Formal Ontology and Information Systems FOIS-2004, 2004.
10. *Александрова З. Е.* Словарь синонимов русского языка // М.: Русский язык, 1999.
11. *Лукашевич Н. В., Добров Б. В.* Тезаурус русского языка для автоматической обработки больших текстовых коллекций // Компьютерная лингвистика и интеллектуальные технологии: Труды Международного семинара Диалог’2002 / Под ред. А. С. Нариньяни. М.: Наука: 2002. Т. 2. С. 338–346.
12. *НОСС.* Новый объяснительный словарь синонимов русского языка. Третий выпуск. Под общим руководством акад. Ю. Д. Апресяна. М.: Языки славянской культуры, 2003.

Анализ текстов SMS-сообщений с целью повышения качества их автоматического озвучивания

Analysis of SMS texts for better automatic reading of sms messages

Людовик Т. В. (tetyana.lyudovyk@gmail.com)

Международный научно-учебный центр информационных технологий и систем НАН Украины и МОН Украины, Киев, Украина

Статья посвящена выявлению особенностей SMS-текстов, препятствующих качественному автоматическому озвучиванию SMS-сообщений. Основные проблемы связаны с неправильным распознаванием языка SMS-сообщения, использованием при написании SMS суржика и сленгов, нестандартной транслитерации и орфографии.

1. Введение

Решение актуальных задач автоматического распознавания и синтеза речи связано с учетом особенностей языковой ситуации, сложившейся в сферах применения этих речевых технологий. Задачи усложняются, если приложения связаны с живым разговорным языком.

Большинство разработанных систем синтеза речи [1–4] озвучивают правильные с точки зрения орфографии и синтаксиса тексты на литературных языках. В условиях многоязычности возникает проблема распознавания языка, на котором написаны тексты.

В данной статье анализируется опыт использования системы синтеза украинской речи для озвучивания SMS-сообщений, отправляемых с мобильных телефонов на стационарные [5]. Большинство текстов SMS-сообщений являются спонтанными, и их автоматическое озвучивание требует учета не только лингвистических, но и социальных, демографических и психологических факторов.

Помимо спонтанности тексты SMS-сообщений отличаются несоблюдением норм литературного языки, использованием сленгов с нефиксированным составом словарей и экспрессивностью. Задачу озвучивания SMS-сообщений в Украине осложняет также сложившаяся языковая ситуация, характеризующаяся двуязычием и наличием смешанного украинско-русского языка — суржика.

По данным переписи населения Украины 2001 года этнический состав страны, определявшийся моноэтнической самоидентификацией взрослых и собственной идентификацией ими своих детей, насчитывал 77,8 % украинцев, 17,3 % русских и 4,9 % представителей других этнических групп. При этом украинский язык назвали родным 67,5 % населения Украины, а русский — 29,6 % украинских граждан (по данным Госкомстата Украины [6]).

В рамках всеукраинских социологических опросов, проведенных Киевским международным институтом социологии в 1991–2003 годах (около 173 000 интервью) [7] проводились устные интервью, целью которых было определить, на каких языках общается взрослое население (старше 18 лет). На вопрос о том, какой язык является наиболее удобным при общении, 47 % граждан Украины назвали украинский и 53 % — русский.

Однако, самоидентификация языка респондентами не отражает их фактического речевого поведения. Проведенные Киевским международным институтом социологии опросы ценны лингвистическими экспертными оценками, данными по окончании интервью интервьюерами-билингвами. Эти оценки свидетельствуют о том, что довольно значительная часть респондентов разговаривала с интервьюерами на смеси украинского и русского языков. Так, в 2000–2003 годах 38,7 % взрослого населения Украины говорило на украинском языке, 46,6 % — на русском и 14,7 % — на суржике.

Часто во время проведения социологических опросов суржикоязычное население фиксируется как украиноязычное. Данные приведенных опросов косвенно это подтверждают: использование суржика зафиксировано у 14 % этнических украинцев и у 5 % этнических русских. Таким образом, озвучивание SMS-сообщений, написанных на суржике, должно осуществляться системой синтеза украинской речи.

2. Цель работы

Целью работы является исследование и учет особенностей SMS-сообщений для более качественного их озвучивания системой синтеза речи. Необходимо классифицировать выявленные проблемы, определить наиболее важные из них с точки зрения влияния на качество синтезированной речи, и выяснить, возможно ли их решение на данном этапе. Для проблем, поддающихся решению, необходимо предложить пути решения.

3. Исследованный материал

Материалом для проведенных исследований послужили реальные SMS-сообщения, отправленные с мобильных телефонов в периоды с 17 мая по 6 июня и с 19 по 27 июля 2008 года. Эти SMS-сообщения были анонимизированы путем исключения данных об отправителях и получателях. Таким образом, анализировались лишь тексты сообщений. Экспертами было выделено пять категорий SMS-текстов в соответствии с языком, на котором они написаны. В таблице 1 представлено распределение проанализированных SMS-сообщений. К категории «условно-украинский язык» отнесены:

- SMS, язык которых невозможно однозначно определить, поскольку все слова текста принадлежат как русскому, так и украинскому языкам (хотя, возможно, произносятся по-разному (ср. «напиши» ([нап'ишы] и [напышы]), «день ([д'эн'] и [дэн']));
- SMS, написанные на суржике.

В данной работе использовались также результаты тестирования компонента автоматического распознавания языков сервиса SMS2Voice [5]. В рамках этого сервиса, работающего в Украине, все SMS-сообщения, отправляемые с мобильных телефонов на стационарные, на первом этапе обработки автоматически классифицируются по языку. В список возможных языков включены только русский, украинский и английский; в случае, если язык определить не удается, по умолчанию языком SMS-сообщения считается русский. После распознавания языка сообщения оно озвучивается соответствующей моноязычной системой синтеза речи.

Автоматическое распознавание языка 5480 SMS-сообщений дало следующие результаты: на русском языке написано 73 % SMS-сообщений, на украинском — 14 %, на английском — 4 %. В 10 % случаев язык определить не удалось.

Таблица 1. Распределение SMS-сообщений по языковым группам

Язык	Количество SMS (17.05.2008–6.06.2008)	Количество SMS (19.07.2008–27.07.2008)	Примеры
Русский	7640 (70,6 %)	3746 (68,4 %)	*Я на работе.*
Украинский	1523 (14,1 %)	843 (15,4 %)	*Як справи? (Как дела?)*
Условно-украинский	1262 (11,7 %)	586 (10,7 %)	*Напиши SMS.* *Умееш находить похидну (производную) с корня?*
Другие языки и нетекстовые SMS	400 (3,7 %)	305 (5,6 %)	*Tutto ok. E mille grazie @1@2@3@*
Всего	10 825 (100 %)	5480 (100 %)	

Таблица 2. Сравнение результатов распознавания языка SMS-сообщений

Язык	Автоматическое распознавание	Экспертное распознавание
Русский	3998	3516
Украинский	748	665
Английский	200	138
Язык не определен	534	
Условно-украинский		945
Другие языки и нетекстовые SMS		216
Всего	5480	5480

4. Многоязычность текстов SMS-сообщений

Проведенное сравнение результатов автоматического распознавания языка SMS-сообщений с результатами распознавания языка этих же SMS-сообщений экспертами-лингвистами дало результаты, приведенные в таблице 2.

Эксперты отнесли к русскому языку существенно меньшее количество SMS-сообщений. Это может быть объяснено тем, что автоматическое распознавание связано с поиском слов в словарях, и если слова текста SMS обнаруживаются как в русском, так и в украинском словарях, то автоматически принимается решение о русском языке. Однако сочетание обнаруженных слов, учитываемое экспертами, может однозначно свидетельствовать в пользу украинского языка. Например: «*я тебе люблю!*». В дальнейшем при распознавании языка SMS-сообщения необходимо учитывать не только наличие слов в словарях, но и их частотность в каждом из языков, ср.: «*як там?*» («*как там?*»).

Существенно меньшее количество SMS, распознанных экспертами как англоязычные, связано с тем, что не учитываются другие неславянские языки. Среди других языков экспертами были выявлены следующие: французский, немецкий, испанский, итальянский, турецкий, вьетнамский, нидерландский, молдавский/румынский, венгерский и грузинский. В настоящее время вопрос озвучивания SMS на этих языках не решен.

В целом, нетранслитерированные орфографически правильно написанные SMS-сообщения на украинском, русском и английском языках не представляют значительных трудностей при их озвучивании.

5. Особенности SMS-сообщений, написанных на украинском и условно-украинском языках

В дальнейшем проводился анализ текстов SMS, отнесенных к украинскому и условно-украинскому языкам. Из категории «условно-украинский язык» были исключены тексты, которые могут быть отнесены к русскому языку, например, «*6 вагон*.», «*Я буду завтра*.». Поскольку SMS на суржике должны озвучиваться системой синтеза украинской речи, украинские и условно-украинские SMS были объединены в одну категорию «SMS на украинском языке». Всего в дальнейшем анализировалось 1529 SMS-сообщений.

Наличие словаря литературного языка, словаря ненормативной лексики, словаря молодежного и SMS-сленгов, а также таблиц транслитерации позволяет использовать для разграничения SMS-текстов критерии, перечисленные в таблице 3. Наиболее сложно разграничить SMS, написанные на литературном украинском языке с орфографическими ошибками (описками), и SMS, написанные на суржике.

К группе «SMS на литературном языке с нестандартной орфографией» отнесены SMS с сокращениями слов.

В таблице 4 приведено распределение SMS-сообщений по группам.

Как правило, озвучивание орфографически правильно написанных на литературном языке SMS-сообщений не представляет затруднений.

Таблица 3. Критерии разграничения SMS-текстов

Критерии разграничения	Примеры
Все слова текста SMS-собщения найдены в словаре литературного языка непосредственно или после обратной транслитерации	*Вітаю з днем народження!* *Ja vzhe na misci.*
Прочитанное вслух SMS-собщение с нестандартной орфографией или нестандартной транслитерацией звучит на литературном языке	*ЯНА ЗВАРЫ БОРЩ И ВИДРО ВАРЕНЫКИВ*
Хотя бы одно слово текста SMS входит в список ненормативной лексики	*Бачу тобі …!*
Хотя бы одно слово текста SMS написано на молодежном или SMS-сленге	*Я в унівєрі. Пліз.:-(*
Все остальные SMS считаются написанными на суржике	*Визивай пожарних на склади загорилися ящики.*

Таблица 4. Распределение SMS-сообщений, написанных на украинском и условно-украинском языках

Категории SMS	Количество SMS
SMS, орфографически правильно написанные на литературном языке	663
SMS на литературном языке с нестандартной орфографией и/или транслитерацией	590
SMS с использованием ненормативной лексики	42
SMS с использованием молодежного и/или SMS-сленга	20
SMS на суржике	214
Всего	1529

6. SMS-сообщения с нестандартной / неоднозначной транслитерацией

Значительная часть исследованных SMS-сообщений (17 %) написана с использованием латинского алфавита, что обусловлено экономией времени и денег, а также возможностями кодировок. В более половины случаев игнорируется наличие официальных таблиц транслитерации.

Наиболее часто отклонения происходят при транслитерации йотированных гласных букв и сочетаний «йотированная гласная + "й"» («*lublu*» вместо «*lyublyu*», «*daite*» вместо «*dajte*»). По-разному транслитерируются «ж», «ш», «ц», «щ», «с». Часто при транслитерации исчезают мягкий знак и апостроф. Широко используются цифры 1 и 4, реже — цифры 0, 3, 6.

Справедливости ради следует отметить, что таблицы транслитерации неудобны, и даже мобильные операторы, посылая SMS-сообщения, не соблюдают правила транслитерации.

Часто одна и та же латинская буква соответствует различным украинским; иногда это случается в одном и том же слове, например, «*rozbudulu*» (латинская «u» используется как в качестве «и», так и в качестве «у»).

В настоящее время учет нестандартной транслитерации латиницей осуществляется следующим образом. Для каждого транслитерированного слова порождается множество вариантов записи кириллицей с учетом вероятности замен латинских букв кириллическими. Затем все варианты в порядке убывания вероятности проверяются на наличие в словаре.

Особую группу составляют украинские SMS-сообщения, транслитерированные русскими буквами, например, «*Який ти писля цёго друг.*», «*Я для неї вірш напысав.*», «*А ты де на Днипри?*». Подобные SMS либо распознаются как русскоязычные и озвучиваются системой синтеза русской речи, либо распознаются как украиноязычные и озвучиваются по-украински с ошибками.

7. SMS с нестандартной орфографией

Нарушение норм пунктуации при написании SMS-сообщения в данной работе не рассматривается, поскольку оно в значительно меньшей степени влияет на качество озвучивания, чем нарушение норм орфографии. Нестандартная орфография зафиксирована в 39 % украинских SMS (не считая SMS, написанных на суржике), что представляет собой одну из главных проблем при озвучивании.

Отклонения от стандартной орфографии могут быть намеренными («*Яаа лююблю тильки тебеее!*»), в результате описок («*на дороагах*») и неграмотности («*Візьми тіліфон.*»).

Несмотря на то, что при опросах 73 % респондентов-украинцев заявили, что хорошо владеют письменным стилем украинского языка, и 71 % — устно-разговорным, самооценка не соответствует реальному уровню владения языком. Не может не тревожить неграмотность молодежи.

Некоторые часто встречающиеся в текстах SMS орфографически неверно написанные слова и словосочетания внесены в словарь и в настоящее время озвучиваются правильно, например, «*будь ласка*», «*будь-ласка*», «*будьласка*». Необходим более планомерный подход к проверке орфографии с учетом того, что часто соседние слова пишутся слитно, а одно слово может быть разделено на части.

8. SMS с нестандартными сокращениями

Нестандартные сокращения в SMS-текстах не могут быть расшифрованы. Как правило, по тексту, состоящему из одного-двух предложений, сложно восстановить семантический контекст. Неясно, когда системы синтеза речи будут в состоянии определить, что в тексте «*Вона вже говорила, шо ти весь час пропускаеш к. р.*» речь идет о школьных контрольных работах, а в тексте «*З ДН тебе*» — о дне рождения.

9. SMS с использованием элементов сленгов

Молодежный и SMS сленги отличаются динамичностью, постоянным появлением новых слов и выражений. Очевидно, для озвучивания соответствующих SMS-сообщений необходимы специальные словари.

Другой отличительной особенностью этой группы SMS-сообщений является экспрессивный характер. Озвучивание SMS-приколов, возможно, требует особой выразительной просодики. Это должно быть учтено на этапе разработки речевой базы данных при составлении текстов, на основе которых производятся акустические записи голоса диктора. Диктор, чей голос впоследствии будет использован для озвучивания SMS, должен читать эти тексты с подчеркнутой выразительностью.

10. SMS с использованием ненормативной лексики

Существует список слов ненормативной лексики. При озвучивании слова из этого списка заменяются сигналом «би-и-п». К сожалению, зафиксировать список не удается. Некоторые пользователи упражняются до тех пор, пока, видоизменяя неприличные слова, не добьются желаемого звучания.

11. SMS на суржике

Как показывают результаты социолингвистического мониторинга, на суржике общаются 15 % взрослого населения Украины и 27 % студентов.

В данной работе принято рабочее определение суржика как смеси украинского и русского языков. Для удобства анализа разделение украиноязычных SMS-сообщений произведено на непересекающиеся группы. Группа «SMS на суржике» содержит только те сообщения, которые не вошли в остальные группы (см. таблицы 3 и 4). В действительности, элементы суржика могут присутствовать в одном SMS-сообщении наряду с ненормативной лексикой и молодежным сленгом. По нашим данным, на суржике пишется от 14 % до 18 % украиноязычных SMS-сообщений.

Как правило, лексика в суржике взята из русского языка, а большая часть грамматики — из украинского. Простые случаи, когда русские слова встречаются в текстах на суржике в неизменном виде («*срочно*», «*тоже*»), могут быть учтены с помощью расширения украинского словаря. Дополнительное неправильное орфографическое написание («*нада*», «*пожалуста*») усложняет задачу.

Более сложные случаи взаимопроникновения языков («*відказуємося*», «*задержуюсь*») нуждаются в дополнительном анализе. Существует мнение, что суржик — это индивидуальное нарушение языковых норм, что он всегда персонален и, следовательно, не может быть закреплен в словарях и грамматиках. Однако, в соответствии с другой точкой зрения, суржик возникает в результате системной интерференции. В настоящее время суржик в основном рассматривается как испорченный русизмами украинский язык. Возможно, в будущем он получит иной статус и будет составлен словарь суржика. Аналогично, будут разработаны системы синтеза речи на суржике.

12. Выводы

Языковая ситуация в Украине изучена недостаточно. Фиксация суржикоязычного населения как украиноязычного не отражает реального соотношения между языками общения.

Учет прагматических факторов, влияющих на написание SMS сообщений, способствует более точному автоматическому распознаванию языка сообщения.

Основные проблемы, возникающие при озвучивании текстов SMS, связаны с нестандартными транслитерацией и орфографией, использованием суржика и сленгов.

Для озвучивания SMS, написанных на распространяющемся быстрыми темпами суржике, необходимо либо расширение речевых баз данных общеукраинского языка, либо создание отдельных, специальных речевых баз данных суржика. Последнее не исключено, если суржик будет признан самостоятельным языком общения.

Литература

1. *Quazza S., Donetti L., Moisa L., Salza P. L.* ACTOR: a Multilingual Unit-Selection Speech Synthesis System // 4th ISCA Tutorial and Research Workshop on Speech Synthesis. 2001. Paper 209.
2. *Лобанов Б. М., Цирульник Л. И.* Компьютерный синтез и клонирование речи // Минск: Белорус. наука, 2008. 337 с.
3. *Oparin I., Talanov A.* Outline of a New Hybrid Russian TTS System // Proc. of the 12th International conference on Speech and Computer SPECOM 2007. Moscow.: 2007. Pp. 603–608.
4. *Romsdorfer H., Pfister B.* Text analysis and language identifcation for polyglot text-to-speech synthesis // Speech Communication, 2007. Vol. 49, pp. 697–724.
5. *Lyudovyk T., Brozinski S., Noner M., Robeiko V., Sazhok M.* Speech Synthesis Applied to SMS reading // Proc. of the 13th International Conference «Speech and Computer: SPECOM'2009». — St. Petersburg: 2009. Pp. 300–305.
6. http://www.ukrcensus.gov.ua/results/general/language/
7. *Хмелько В. Є.* Лінгво-етнічна структура України: регіональні особливості і тенденції змін за роки незалежності // Наукові записки НаУКМА. «Соціологічні науки». К.: 2004. Т. 32. С. 3–15.

Оценка методов автоматического анализа текста: морфологические парсеры русского языка

NLP evaluation: russian morphological parsers

Ляшевская О. Н., Астафьева И., Бонч-Осмоловская А., Гарейшина А., Гришина Ю., Дьячков В., Ионов М., Королева А., Кудринский М., Литягина А., Лучина Е., Сидорова Е., Толдова С.

МГУ им. М. В. Ломоносова

Савчук С.

ИРЯ РАН им. В. В. Виноградова

Коваль С. (lingtecheval@yahoo.com)

Форум «Оценка методов АОТ» (http://ru-eval.ru) — новая инициатива, целью которой является независимая оценка методов и алгоритмов работы русскоязычных лингвистических ресурсов. В статье описываются принципы и процедура проведения дорожек форума, состав участников, тестовая коллекция, организация экспертизы и полученные результаты.

1. Введение

Форум «Оценка методов автоматического анализа текста» стартовал в феврале 2010 года, и темой первого года стали морфологические парсеры русского языка. Тестовый запуск систем и экспертиза ответов состоялись в марте-апреле, а очную встречу участников и обсуждение результатов предполагается провести на конференции «Диалог'2010». Сама идея форума возникла два года назад на конференции Language Resources and Evaluation (LREC'2008), но настоящим своим рождением она обязана конференции «Диалог». Вместе с постоянными участниками «Диалога» мы обсуждали, отчего, несмотря на существование старых и хорошо зарекомендовавших себя парсеров русского языка, все время появляются новые процессоры, нужны ли лингвистические ресурсы, например, словари, для построения компьютерных лингвистических систем, в чем задача лингвистов на разных этапах развития IT–технологий и, наконец, почему в мире большой популярностью пользуются некоммерческие семинары по сравнительной оценке парсеров (ср. проекты CLEF, AMALGAM, GRACE, EVALITA, SEMEVAL и др.) и не нужно ли ввести такую моду в России для русскоязычных ресурсов.

Ключевое событие форума строится в игровой форме — системы соревнуются друг с другом на специально подготовленной коллекции текстов, кто даст больше правильных ответов. Однако цель соревнования вовсе не в том, чтобы назвать победителя, а в том, чтобы выявить, какие алгоритмы и ресурсы позволяют улучшить результаты по тому или иному показателю. В связи с этим форум предполагается проводить регулярно, чтобы дать разработчикам возможность из года в год совершенствовать свои методы. Таким образом, настоящая высокая цель форума — улучшение состояния науки в области автоматической обработки текста. Но главное, форум должен способствовать созданию среды, в которой научные, научно-производственные, коммерческие разработки могли бы проходить независимую экспертизу, и в которой могли бы обсуждаться проблемы и перспективы развития технологий.

Немаловажным представляется и практический выход, полученный по окончании данного соревнования: корпус вручную размеченных и выверенных текстов, который можно использовать в научно-исследовательских целях, сформированные принципы разметки, к которой могут быть приведены разметки большинства систем, исчисление сложных случаев русского языка, которые не име-

ют однозначного решения. Счастливым образом, форум 2010 года получил также образовательную составляющую — в его подготовке, проведении и формировании финального отчета активное участие принимали студенты Отделения теоретической и прикладной лингвистики филологического факультета МГУ им. М.В.Ломоносова, которые получили возможность «пощупать руками», как работают парсеры, увидеть, в чем их сильные и слабые стороны, чем парсеры системно отличаются друг от друга и т. д.

Объектом рассмотрения в данном форуме являются не собственно морфоанализаторы, работающие с изолированными словами (именно они рассматривались в качестве объекта оценки в отдельных работах последнего десятилетия, ср. Коваль 2003), а модули, учитывающие или потенциально учитывающие контекст. В связи с этим как в названии форума, так и во всей его внутренней документации последовательно используется понятие «морфологический парсер», обозначающее модуль, функциональность которого позволяет, как минимум, обрабатывать сразу всю текстовую цепь слов, и как максимум, учитывать при анализе каждого текстового слова результаты разбора его соседей. В этой второй, «сильной» интерпретации термин «морфологический парсер» становится практически неотличимым от используемого в англоязычной литературе «POS tagger», однако организаторы форума предпочитают говорить о «морфологических парсерах» в силу специфики русского языка: как «слабые» (не предусматривающие контекстную дизамбигуацию разборов), так и «сильные» (включающие такую дизамбигуацию) варианты парсеров опираются на заложенную их разработчиками модель такого далеко не тривиального объекта, как русская словоизменительная морфология, а, значит, имеют достаточно много общего.

Важнейшая презумпция организации соревнования состояла в том, что не бывает единственно правильного решения грамматически спорных вопросов и единственно правильного алгоритма морфологического анализа. Существует множество примеров того, как оптимальный выбор того или иного решения зависит от той цели, для которой проводится анализ. Так, выделение устойчивых словосочетаний как одной единицы (например, «Государственная Дума») может улучшить качество информационного поиска, двукомпонентный анализ, в данном случае, необходим для корректных последующих уровней обработки. Разбор словоформы «бело-кремовое» как единого целого, получающего грамматическую характеристику по концовке, вполне удовлетворителен во многих ситуациях, однако для тех систем, в цикл обработки которых включен семантический анализ, для осмысления этой явно несловарной формы наверняка потребуется ее сегментация по дефису.

В связи с этим достаточно широкий круг грамматических вопросов был вынесен за скобки соревнования и не оценивался. Тем не менее, именно эти проблемы — и расхождения систем в предлагаемых решениях — явились предметом особого внимания со стороны организаторов. Нам представляется, что исчисление и классификация случаев, сложных для автоматического грамматического разбора, а также сведения о частотности возможных решений являются самоценной информацией, которая может быть использована научным сообществом и для исследовательских целей, и для улучшения эффективности прикладных разработок.

2. Дорожки

Организационно форум 2010 года во многом строился по образцу Семинара по оценке методов информационного поиска РОМИП (РОМИП 2009). Оценка алгоритмов проводилась по нескольким независимым дисциплинам (дорожкам). Каждая дорожка была посвящена одной конкретной задаче анализа текста с заранее согласованными правилами оценки систем-участников. От участников не требовалось участия во всех дорожках сразу, поэтому у них была возможность сосредоточиться на решении только одной из предлагаемых задач.

В соревнованиях рассматривались два типа морфологических разборов:

1) без дизамбигуации: системы дают множество возможных разборов, оценивается наличие среди них правильного разбора;
2) с дизамбигуацией: система должна дать единственный правильный разбор, корректность которого является объектом оценки.

Соревнования без дизамбигуации состоялись на следующих дорожках:

- «Лемматизация». Задача этой дорожки состояла в том, чтобы правильно определить исходную форму словоформы
- «POS». Требовалось правильно определить часть речи, к которой принадлежит исходная словоформа
- «Морфология». Задача: правильно определить грамматические теги, которые характеризуют исходную словоформу, например, род, число, падеж, время и т. д. Оценивалось наличие правильной комбинации грамматических тегов, представленных в разборе
- «Редкие слова». Задача состояла в том, чтобы правильно определить лемму и часть речи для списка специально отобранных несловарных или нестандартных словоформ.

Соревнования с дизамбигуацией проводились на дорожках:

- «Дизамбигуация: леммы» и
- «Дизамбигуация: POS».[1]

3. Участники

На конкурс были поданы заявки от 15 групп разработчиков из Москвы, Санкт-Петербурга, Екатеринбурга (Россия), Минска (Беларусь), Донецка (Украина), Лидса (Великобритания). В тестовых дорожках приняли участие 12 систем: ARME, Crosslator, FSTMorph (+ ЭТАП-3), Libmorphrus, Mocky, Mystem (+ FastDictionary), Polymorph, Pymorphy, RDMA_IAI, Semantarus Morpho, Starling, TextAn[2]. Некоторые разработчики представляли несколько вариантов морфологических анализаторов для дорожек с дизамбигуацией и без нее и даже несколько вариантов реализации алгоритмов на одной дорожке.

В итоге было получено 13 ответов систем по дорожкам «Лемматизация» и «POS», 12 ответов по дорожке «Морфология», 8 ответов по дорожке «Редкие слова» и 7 ответов по обеим дорожкам с дизамбигуацией. Ответы одного участника по дорожкам «Лемматизация», «POS» и «Морфология» были дисквалифицированы за несоответствие формата данных и не участвовали в экспертизе.

4. Тестовая коллекция и задания

Для соревнования была подготовлена общая коллекция неразмеченных текстов для дорожек «Лемматизация», «POS», «Морфология», «Дизамбигуация: леммы» и «Дизамбигуация: POS» (Основная коллекция) объемом около 1 млн. словоупотреблений. Материалы для Основной коллекции были составлены из фрагментов текстов, присланных некоторыми участниками и экспертами. В Основную коллекцию вошли тексты различной тематики и жанровой принадлежности в следующих соотношениях: 18 % Статьи в СМИ/Нонфикшн, 15 % Новости; 15 % Интервью; 15 % Технические тексты; 15 % Юридические тексты; 18 % Художественная литература; 4 % Блоги и форумы.

На базе Основной коллекции было составлено задание для дорожки «Редкие слова», включавшее 75 отобранных экспертами слов с их ближайшим контекстом, в том числе:

1) продуктивные модели (слова с неизвестным словарю корнем, но образованные с помощью продуктивных аффиксов. Среди них встречаются так называемые слова-обманки: *аррабьята* (лемма «аррабьята») vs *френдята* (лемма «френденок») и т. п., а также авторские «придуманные» слова: *увазила, кругтелся, склипких, грезитвой*;
2) сложные слова, у которых вторая часть совпадает со словами или вторыми частями сложных слов в словаре Зализняка: *полуколебаний*, *ультраженственной, миллионометра, Росторгмонтаж*;
3) слова с «неизвестными» корнями (в т. ч. имена собственные), не содержащие продуктивных аффиксов, для которых носители языка могут однозначно определить лемму и часть речи (по стандартным окончаниям русского языка и зная контексты, в которых они употреблены): *турбийона* (лемма «турбийон»), *френдя* («френдить»), *тюрбо* («тюрбо»), *Баухаус* («Баухаус») и др.;
4) редкие и нестандартные формы (некоторые деепричастия, формы первого лица глаголов и степени сравнения, которые употребляются в языке, но признаются окказиональными или ненормативными, в связи с чем обычно отсутствуют в словарях): *стригя, пья, побежу*, *висю*, *деревяннее, нельзей*;
5) Аббревиатуры типа *ВЧК*, *ОГПУ, МФТИ*, которые система могла бы спутать с глаголами или словами других классов и ошибиться в определении леммы.

Источником выборки редких слов послужили научные тексты, инструкции, кулинарные рецепты и меню, записи речи детей дошкольного возраста (большинство интересных продуктивных моделей и нестандартных форм было обнаружено именно там, поскольку в возрасте с 3 до 5 лет дети постоянно изобретают новые слова), форумы в Интернете, а также тексты Велимира Хлебникова и Людмилы Петрушевской. Итоговый баланс задания «Редкие слова» включает 27 существительных, 13 прилагательных, 28 глаголов и 7 слов категории ADV.

Сравнение результатов по всем дорожкам проводилось на основе выборочной проверки ответов систем-участников. Для этого был подготовлен «Золотой Стандарт» — множество случайно выбранных предложений из Основной коллекции, объемом около 2000 словоупотреблений. В ходе экспертизы ответы систем сравнивались с произведенной экспертами ручной разметкой Золотого Стандарта, см. п. 6.

---

[1] Первоначально предполагалось также проведение дорожки «Коллекции: Грязные тексты», где системам ставилась задача разметить фрагменты плохо распознанных отсканированных документов, таблиц, содержащих слова с некорректно внесенными знаками переносов и форматирования и текстов с большим количеством опечаток. Была подготовлена и разослана участникам специальная коллекция, однако, поскольку по этой дорожке был получен только один ответ, дорожка была отменена и экспертиза результатов по ней не проводилась.

[2] Еще одна система (АОТ) выступала вне зачета, с согласия автора ее запускали студенты-эксперты. Более подробную информацию об участниках можно найти на странице http://ru-eval.ru/participants.html.

5. Принятые соглашения по унификации грамматической информации

Подготовительный этап потребовал определенных решений, направленных на унификацию нотации и структуры морфологических разборов в ответах, ожидаемых от парсеров. Было выявлено несколько типов проблемных случаев:

1) некоторые частеречные категории не имеют устойчивой общепринятой нотации разметки и выделяются, обозначаются и объединяются системами по-разному, что может затруднить оценку результатов (например, в одних системах выделяется один общий класс местоимений, в других системах они разводятся по классам существительных, прилагательных, наречий и т. д., в третьем случае выделяются классы местоимений-существительных, местоимений-прилагательных и т. п.);
2) объем парадигмы может разниться от системы к системе, например, формы парных глаголов совершенного и несовершенного вида могут приводиться к двум разным леммам (*прыгнул — прыгнуть*, *прыгал — прыгать*) или к одной общей (*прыгать*); часто само требование к объему парадигмы зависит от того, для решения какой прикладной задачи используется модуль морфологического парсинга;
3) некоторые классифицирующие признаки словоформ (например, переходность у глаголов) могут считаться избыточными на этапе морфологического анализа текста, а их определение может быть затруднено в том случае, если анализируемая словоформа не входит в словарь системы;
4) некоторые морфологические признаки не могут быть однозначно определены в рамках морфологического анализа (например, нетривиально определение леммы и залога для глаголов с постфиксом *-ся*);
5) некоторые морфологические характеристики (например, звательный падеж) имеются только у ограниченного числа словоформ и могут системно не выделяться.

С учетом ожидаемых расхождений было принято решение о том, что разметка будет производиться парсерами по упрощенной системе. При лемматизации буквы *е* и *ё*, а также написание с прописной/строчной буквы признавались равноправными. Частеречные признаки были приведены к следующему сокращенному инвентарю: существительные (S), прилагательные (A), глаголы, в том числе причастия и деепричастия (V), предлоги (PR), союзы (CONJ), и сборная категория, включающая прочие несклоняемые слова: наречия, вводные слова, частицы, междометия (ADV). Не участвовали в оценке и могли быть размечены любым образом местоимения (включая наречные и предикативные), числительные, а также составные предлоги и союзы (ср. *потому что, в течение*).

Кроме того, был сокращен и список грамматических характеристик, приписываемых словоформе. В общем случае, сопутствующий набор грамматических признаков определялся тем минимумом информации, который нужно знать для однозначного восстановления словоформы из леммы. Морфологические признаки указывались только для существительных, глаголов и прилагательных.

Итоговый список размечаемых морфологических характеристик словоформ включает:

- род: m (мужской), f (женский), n — (средний)
- падеж: nom (именительный), gen (родительный, в том числе счетная форма — два шар/а), dat (дательный), acc (винительный), ins (творительный), loc (предложный, в том числе второй предложный, ср. *в лесу*)
- число: sg (единственное), pl (множественное)
- время: pres (= непрошедшее: настоящее и будущее время — *пишу, напишу*), past (прошедшее),
- наклонение: imper (повелительное)
- инфинитив: inf
- причастие: partcp,
- деепричастие: ger
- залог: act (действительный), pass (страдательный) — указывается только в формах причастий
- лицо: 1p, 2p, 3p

Таким образом, из классифицирующих категорий необходимым для указания являлся только род, не рассматривались переходность и вид глагола, залог для всех форм глагола кроме причастий и деепричастий, одушевленность имен. Кроме того, необязательно было указывать при разборе степень сравнения прилагательных и наречий, а также полноту/краткость прилагательных.

Следует также отметить, что не участвовал в оценке целый ряд непродуктивных словоизменительных категорий, а также маргинальных реализаций продуктивных категорий: лицо и наклонение форм императива 1 лица типа *пойдемте*; падеж имен в конструкциях «пойти в *солдаты*», «попить *чаю*»; звательный падеж (*Маш! отче* и др.); род слов общего рода (*врач*).

6. Подготовка Золотого Стандарта

Ручная разметка Золотого Стандарта, предшествовавшая экспертизе результатов, преследовала несколько целей. Во-первых, требовалось независимое основание для автоматического сопоставления ответов систем, которое уменьшило бы объем ручной экспертизы: проверке подлежали только случаи

расхождения между стандартом и ответами систем. Во-вторых, организаторы хотели избежать влияния результатов, предоставленных системой, на интуицию экспертов, и пропусков ошибок по невнимательности. В–третьих, разметка Стандарта должна была подготовить экспертов к оценке ответов систем, сформировать у них представление о том, какие сложные случаи их ожидают, понять объективную природу несовпадения некоторых ответов и выработать критерии для их либеральной оценки.

В разметке Стандарта принимало участие 10 экспертов, каждый фрагмент размечался независимо двумя разметчиками. Перед ними стояла задача выделить в тексте все русские словоформы и дать им единственный разбор. После технической валидации разметки на предмет соблюдения формата и допустимых сочетаний тегов согласованность результатов ручной разметки (inter-annotator agreement) составила: леммы — 94,4 %, POS — 95,4 %, морфология — 89,0 %, весь разбор в целом — 85,5 %. Оставшиеся содержательные расхождения согласовывались экспертами в паре. В случае если эксперты не могли прийти к единому решению, спорные вопросы выносились на обсуждение на специально организованных семинарах с участием всех разметчиков и еще 5 экспертов. В частности, обсуждалось, как лемматизировать потенциальные pluralia tantum, сокращения, слова с дефисом или незнакомые слова; к какому классу принадлежат слова типа *минувший*: причастие или отпричастное прилагательное; *данные*: прилагательное или отадъективное существительное? Каждый эксперт высказывал свое мнение по поводу того или иного случая, а также объяснял свою точку зрения. Затем наиболее убедительное решение вносилось в Золотой Стандарт. Например, в случае выбора леммы для *72-часовых* было предложено три возможных решения: 1) это две словоформы, которым приписываются две леммы: «72» и «часовой»; 2) лемма — «72-часовой»; 3) лемма — «семидесятидвухчасовой». В ходе дискуссии предпочтение было отдано первому варианту, который и был отражен в Золотом Стандарте.

7. Экспертиза ответов систем

Процедура экспертизы ответов морфологических парсеров предусматривала сравнение разбора каждой входящей в зачет словоформы с ее разбором в Золотом Стандарте. Полное совпадение по одному из учитываемых параметров (лемма, часть речи, грамматические признаки) автоматически получало оценку 0. При этом на дорожках без дизамбигуации для признания ответа правильным достаточно было наличия правильного разбора среди любого количества вариантов разбора, предложенных системой.

Случаи расхождений отправлялись на рассмотрение экспертам, которые должны были оценить их по следующей шкале:

1 — права Система;

2 — прав Золотой Стандарт;

3 — спорный грамматический вопрос;

4 — затрудняюсь определить (такие оценки впоследствии пересматривались в более широком кругу экспертов);

5 — неправы оба — и Система, и Стандарт.

Сравнение ответов систем с Золотым Стандартом позволило выделить наиболее распространенные отклонения от разборов, признанных эталонными.

1. Существенную часть ошибок составляет неправильное распознавание нестандартных классов слов. Можно выделить 5 основных типов.
   1.1. Слова, имеющие дефис в графической репрезентации. Многие парсеры последовательно разбивают такие слова на части и лемматизируют их по отдельности, что можно признать правильным лишь в небольшом количестве случаев. Правомерность такого разбиения зависит от статуса элементов, составляющих дефисную конструкцию. Так, первым элементом может быть префиксоид (*штаб-квартира*), первый сегмент заимствований, не несущий в русском языке смысловой нагрузки (*Тянь-Шаня, холд-ап*), неотделимая часть некоторых типов предлогов (*из-за*) и наречий (*по-птичьи*) и т. д., и тогда подобное решение грамматически некорректно. Разбиение наиболее правомерно лишь тогда, когда обе части такого формального слова склоняются (например, когда одна из них является приложением к другой: *шофер-предприниматель*) и первая часть может обладать самостоятельными грамматическими признаками, но эти случаи составляют незначительную долю всех слов с дефисами.
   1.2. Некоторые имена собственные. Неверно распознаются и лемматизируются по исходному сегменту. Проблемы частеречной принадлежности и грамматических признаков возникают не только с экзотическими словами, но и с фамилиями на *-ов*, *-их* и т. п.
   1.3. Аббревиатуры. В отдельных случаях не распознаются вообще, некоторые системы опознают только часть речи, в той или иной мере — грамматические признаки.
   1.4. Редкие слова. Зачастую также не распознаются или лемматизируются путем копирования сегмента исходного текста. Иногда по такой неправильной лемме определяются грамматические признаки.

1.5. Общепринятые сокращения типа *тыс.*, *ст.* («статья») и др.

Таким образом, большая часть ошибок возникает в «несловарных» словах, что объясняется тем, что парсеры либо имеют недостаточно эффективные средства обработки таких слов, либо вовсе их не имеют, полагаясь на закрытый список, составляющий словарь системы. Обилие ошибок с определением части речи и грамматической характеристики таких слов указывает на необходимость использования методов, учитывающих контекст. Экспертиза дорожки «Редкие слова» показала, что наиболее уязвимы для парсинга слова непродуктивных моделей (*джоулево, гильоше*), а также глагольные и наречные словоформы. Как кажется, это связано с тем, что для многих прикладных задач выбор в пользу продуктивных моделей и имен существительных дает большую эффективность системы.

2. Омонимия
   2.1. Достаточно типичными являются ошибки при разборе частичных (не «системных») омонимов, которые могли неверно лемматизироваться *(парный — парной)* и, как следствие, получали неверную POS-характеристику (*ели*).
   2.2. Особый класс среди омонимов составляют пары из глаголов и отглагольных прилагательных/существительных (*окружающий* как форма глагола и как прилагательное, *данные* как форма глагола и как существительное), наречий и прилагательных (*ясно* как форма наречия или прилагательного), а также наречий и производных предлогов (*вблизи, навстречу*), для различения которых нельзя обойтись морфологическими критериями. Это обстоятельство вызвало некоторые колебания среди экспертов в оценке таких случаев.
3. Часть ошибок можно объяснить неправильным разбором по аналогии. Наиболее типичным случаем является ошибочная лемматизация глаголов с постфиксом *-ся* путем отсечения этого постфикса в ситуации, когда соответствующий парный глагол не существует или отчетливо отличается по значению. Например, для глаголов типа *являться*, *стремиться*, *находиться* отдельными системами были предложены в качестве лемм, соответственно, *являть, стремить, находить*.
4. В отдельных случаях участники использовали классификации частей речи, которые не совпадали с предварительно заданной для данного соревнования, а потому использование символов этих классификаций оценивалось как ошибочное. Вместе с тем, по общей договоренности, исключение было сделано для числительных и местоимений, разбор которых не входил в зачет.

Наряду с вышеперечисленными типовыми ошибками был выделен ряд случаев лемматизации, определения части речи и полного грамматического разбора, которые по общему мнению были квалифицированы как спорные (оценка 3) и допускали более одного правильного (не наказываемого штрафными баллами) варианта. Основные спорные грамматические вопросы включали:

1) определение леммы сравнительных и превосходных степеней наречий и прилагательных (показатель степени может сохраняться в лемме, или же может быть использована лемма положительной степени[3]);
2) определение леммы краткой формы прилагательного (лемматизация по полной / краткой форме);
3) определение леммы парных по виду глаголов (лемматизация по несовершенному виду / по совершенному виду / по тому виду, который присутствует в исходной словоформе);
4) определение леммы глагольных словоформ с постфиксом *-ся* (лемматизация с сохранением постфикса / без него[4]

8. Результаты соревнования

В основу ранжирования ответов систем положены три базовые величины:

- n, общее количество ответов на дорожке — принято за константу для всех систем и соответствует числу словоформ, получивших разметку в Золотом Стандарте и входящих в зачет в соответствии с регламентом;
- f, количество неправильных ответов системы на дорожке: неправильными считаются ответы, получившие оценку экспертов 2 и 5 (см. выше п. 7);
- t, количество правильных ответов системы на дорожке: правильными считаются ответы, получившие оценку 0, 1, 3 и 4.

Организаторы форума не могли уступить искушению использовать такие популярные метрики качества функционирования лингвистических информационных систем, как точность и полноту. Вместе с тем при более внимательном рассмотрении выяснилось, что эти метрики могут быть использованы

[3] Во втором случае формы наречий должны быть приведены к наречиям, а формы прилагательных к прилагательным.

[4] В последнем случае имеется в виду страдательный залог невозвратного глагола. Варианты лемматизации признаются равноправными за исключением тех случаев, когда глагол не употребляется без *-ся* (*удаваться — *удавать*) или же значение глагола без *-ся* принципиальным образом отличается от значения возвратного глагола (*находить — находиться*)

лишь в весьма усеченном виде, по крайней мере на начальном этапе существования форума, когда все процедуры, в том числе оценочные, только отрабатываются.

Это несоответствие связано с принципиальными отличиями в функциональной архитектуре между информационным поиском, из которого берут начало точность и полнота, и морфологическим парсингом. В ситуации оценки информационного поиска все пространство используемой коллекции документов делится на четыре области:

- t_p — документы, признанные релевантными и найденные тестируемой системой,
- f_n — документы, признанные релевантными и не найденные тестируемой системой,
- f_p — документы, не признанные релевантными, но найденные системой,
- $(n — (t_p + f_n + f_p))$ — все остальные документы,

что позволяет определить точность Precision как отношение $t_p/(t_p+f_p)$, а полноту Recall как отношение $t_p/(t_p+f_n)$ и дать этим величинам вполне осмысленную интерпретацию. Однако эта ситуация не находит прямых соответствий в морфологическом анализе текста. Если принять за единицу подсчетов словоформу (а не, допустим, отдельный тег или вариант разбора), то пространство размеченной коллекции текстовых словоформ будет разделено на три области:

- t_p — словоформы, оценка которых учитывается при ответах системы и для которых система дала правильный ответ (= t),
- f_p — словоформы, оценка которых учитывается при ответах системы и для которых система дала неправильный ответ (= f),
- f_n — словоформы, оценка которых учитывается при ответах системы и для которых система не дала ответа (= n − t − f).

Если разбираемый текст содержит словоформы, разбор которых по общей договоренности не подвергается оценке (как местоимения и числительные на данном форуме), случаи их окказионального разбора отдельными системами никак не могут повлиять на оценку этих систем, поскольку остальные участники изначально отказались от их разбора и общее основание для сопоставления результатов всех участников отсутствует. Если одной словоформе из Золотого Стандарта в ответе системы соответствует две словоформы с собственными разборами (например, *бело-кремовое* VS *бело* и *кремовое*), то они получают одну общую оценку. Таким образом, сумма $t_p + f_n + f_p$ является константой (n), обозначающей число словоформ, по которым предполагается давать оценку системе, пользуясь данной версией Золотого Стандарта (это справедливо для всех дорожек — с дизамбигуацией и без дизамбигуации).

Механический перенос формул информационного поиска

$$Precision = t_p/(t_p+f_p)$$

и

$$Recall = t_p/(t_p+f_n)$$

в данную область дает лишь частичный эффект: точность вполне осмысленно характеризует ту пропорцию ответов системы, которой можно доверять, тогда как полнота едва ли может получить разумную интерпретацию. Причиной этому является отсутствие каких-либо общих содержательных признаков для двух слагаемых в знаменателе формулы — числом правильных ответов системы t_p и числом случаев, когда система по ошибке не дала никакого ответа f_n (заметим, что в информационном поиске сумма t_p+f_n давала не что иное, как количество документов, считающихся релевантными для данного запроса). Деление числа правильных ответов на сумму разнородных слагаемых не поддается осмыслению.

Вместе с тем, есть возможность воспользоваться еще одной метрикой, заимствованной из информационного поиска, которой является «аккуратность»:

$$Accuracy = t_p / (t_p + f_n + f_p)$$

В связи с особенностью нашего выбора базовых величин для расчетов (n, f и t) эта метрика имеет вид:

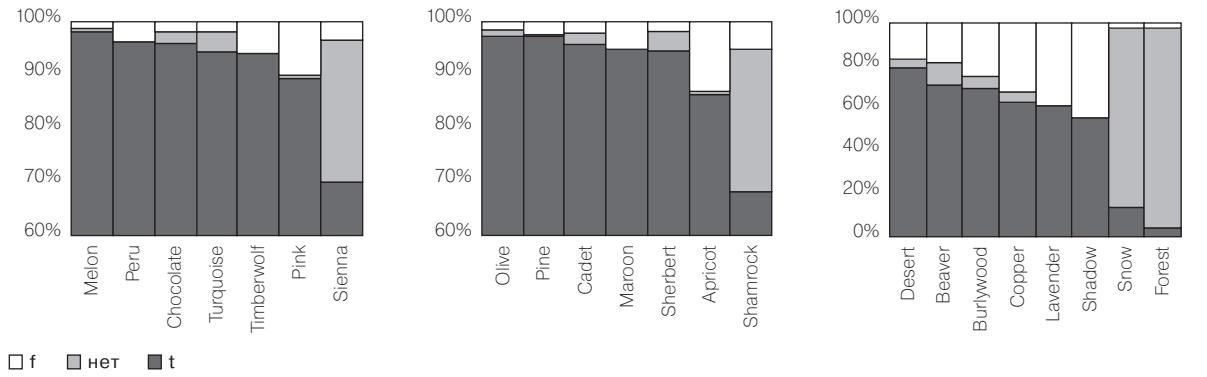
$$Accuracy = t_p / (t_p + f_n + f_p) = t / n$$

и легко интерпретируется как общая оценка качества работы парсера, поскольку позволяет судить о том, какая доля словоформ получит правильный разбор данным парсером.

Существуют иные подходы к определению полноты и точности, см., например, Paroubek 2007: 111–112, где описаны возможные интерпретации этих понятий специально для морфологического анализа без дизамбигуации. При этом либо рассматривается ситуация, допускающая множественность разборов в Золотом Стандарте, что является нетипичным в нашем случае, либо, при сравнении с Золотым Стандартом, приписывающим единственно возможную интерпретацию, полноту предлагается определять так, как у нас определена аккуратность, а точность включает понижающий коэффициент за неразрешенную неоднозначность. Однако мы сочли удобным использовать при экспертизе дорожек 2010 года описанную выше единообразную трактовку метрики для обоих вариантов разбора: как с дизамбигуацией, так и без дизамбигуации.

Таблица 1. Рейтинг систем на дорожках с дизамбигуацией и «Редкие слова»

«Дизамбигуация: Леммы»					«Дизамбигуация: POS»					«Редкие слова»				
Участник	t	нет	f	Accur	Участник	t	нет	f	Accur	Участник	t	нет	f	Accur
Melon	2008	14	24	98,1 %	Olive	1991	22	33	97,3 %	Desert	59	3	13	78,7 %
Peru	1970	1	75	96,3 %	Pine	1991	5	50	97,3 %	Beaver	53	8	14	70,7 %
Chocolate	1964	43	39	96,0 %	Cadet	1958	43	45	95,7 %	Burlywood	52	4	19	69,3 %
Turquoise	1934	75	37	94,5 %	Maroon	1943	0	103	95,0 %	Copper	47	4	24	62,7 %
Timberwolf	1925	0	121	94,1 %	Sherbert	1934	75	37	94,5 %	Lavender	46	0	29	61,3 %
Pink	1831	11	204	89,5 %	Apricot	1769	11	266	86,5 %	Shadow	42	0	33	56,0 %
Sienna	1430	547	69	69,9 %	Shamrock	1394	547	105	68,1 %	Snow	10	63	2	13,3 %
										Forest	3	70	2	4,0 %
Всего ответов				2046					2046	Всего ответов				75
Медиана				94,5 %					95,0 %	Медиана				62,0 %



9. Выводы, перспективы и задачи

Главной целью в 2010 году было положить начало проведению в России семинаров, посвященных оценке методов автоматического лингвистического анализа для русскоязычных коллекций. Как уже отмечалось, в мировой практике сложилась традиция проводить соревнования по различным аспектам автоматической обработки текста, в которых участвуют научные, научно-производственные, коммерческие разработчики, заинтересованные в независимой экспертизе. В России существует такая традиция в области информационного поиска (РОМИП). Однако соревнования, где основное внимание уделяется собственно лингвистическому анализу текста, в русскоязычном сообществе проводилось впервые.

В 2010 году был проведен комплекс работ, в результате которого удалось:

- апробировать организационные процедуры для такого рода соревнования и механизмы взаимодействия, в том числе дистанционного, в рамках оргкомитета;
- собрать большую коллекцию неразмеченных текстов разных жанров, на которой тестировалась работа систем;
- создать коллекцию Золотого Стандарта, размеченную вручную и выверенную несколькими экспертами; эта коллекция может быть использована в дальнейшем для тестирования систем и при подготовке специалистов по прикладной лингвистике;
- выработать основные принципы морфологической разметки для создания Золотого Стандарта;
- принять основные грамматические решения, обеспечивающие унификацию оценки разметок систем;
- выявить сложные и спорные случаи морфологической аннотации, вызывающие затруднения не только при автоматическом анализе, но и при разметке экспертами;
- провести оценку работы парсеров по четырем дорожкам для систем без дизамбигуации и по двум для систем с дизамбигуацией;
- провести содержательный анализ ошибок парсеров, выработать классификацию ошибок систем, а также решений, альтернативных принятым в Золотом Стандарте;
- анализ результатов выявил также сложности в применении к оценке морфологического анализа традиционных метрик, используемых в оценке информационного поиска.

В силу принципиальной несводимости к единому стандарту решений отдельно взятых систем по отношению к ряду спорных вопросов русской морфологии, в 2010 году эти спорные вопросы были вынесены за рамки соревнования. В дальнейшем предполагается постепенно сужать их круг и расширять лингвистическую базу для проведения соревнования, опираясь на взаимодействие с разработчиками морфологических парсеров и учитывая новейшие тенденции в этой области.

Как и ожидалось, анализ результатов работы систем морфологического анализа выявил целый ряд дискуссионных аспектов технологий морфологического анализа:

- состав набора морфологических тегов (специфика категоризации частей речи для различных задач);
- оптимальные соотношения между размером словаря и мощностью генератора гипотез для «несловарных» слов;
- способы борьбы с различными типами «системной» омонимии и др.

Были решены главные задачи форума 2010 года: построение типологии проблем автоматического морфологического анализа текста и оптимизация структурирования соответствующего набора данных, что в целом может служить дополнительным стимулом развития алгоритмов в этой области. Активное участие в соревновании большого количества различных научных и коммерческих коллективов в 2010 г. показало актуальность и востребованность проведения подобных форумов. Проявленный к форуму интерес укрепил уверенность в том, что этот проект положит начало ежегодным соревнованиям, целью которых является оценка методов и алгоритмов лингвистического анализа разного уровня. Последующие мероприятия могут быть посвящены синтаксическому и семантическому анализу, фактографии, анализу звучащей речи, использованию лексикографических ресурсов и многим другим аспектам автоматического анализа текста.

Литература

1. *Коваль С. А.* О сравнимости и эквивалентности компьютерных представлений морфологии // Компьютерная лингвистика и интеллектуальные технологии. Тр. междунар. конференции Диалог'2003 (Протвино, 11–16 июня 2003 г.) / Под ред. И. М. Кобозевой, Н. И. Лауфер, В. П. Селегея. М.: Наука, 2003. С. 305–311.
2. *Российский* семинар по Оценке Методов Информационного Поиска. Труды РОМИП 2009 (Петрозаводск, 16 сентября 2009 г.). Санкт-Петербург: НУ ЦСИ, 2009.
3. *Paroubek P.* On the evaluation of the automatic parsing of natural language // Evaluation of text and speech systems. Text, speech and language technology. Vol. 37. Springer, 2007. P. 99–113.

Генитивная и инструментальная конструкции формы: сходства и различия*

Genitive and Instrumental Constructions of Shape: Similarity and Specificity

Ляшевская О. Н. (olesar@mail.ru)

University of Tromsø (Тромсе, Норвегия)

Рассматриваются две русские конструкции, в которых форма одного объекта характеризуется через форму другого объекта: конструкция с именем в родительном падеже (*иссохшие плети рук*) и конструкция с именем в творительном падеже (*руки повисли плетьми*). Анализируются синтаксические особенности конструкций и взаимозависимости лексического заполнения их слотов. На этом основании делаются выводы о семантических различиях в профилировании пространственных ситуаций.

1. Введение

В русском языке есть несколько конструкций, в которых форма одного объекта метафорически обозначается через форму другого объекта. Здесь мы рассмотрим две из них: генитивную (ср. *иссохшие плети рук, пузырь воздушного шара*) и инструментальную (ср. *руки повисли плетьми*, *раздувшиеся пузырями авоськи*). В обеих конструкциях присутствует имя физического объекта с нетривиальным семантическим признаком «эталона формы» (Сунь Шуан 2009, Гилярова 2002, Кобозева 2000), ср. *плеть, пузырь, груша, змейка, веер, винт, палка, полоса*[1] — далее мы будем обозначать его как *S,* а также имя другого объекта, форма которого характеризуется, ср. *рука* — далее оно будет означаться как *S'*. Важно, что в описываемой ситуации физически присутствует только объект, называемый именем *S'*; груша как плод, змейка как животное или веер как дамский аксессуар в описываемой ситуации не представлены. Иными словами, *груша, змейка* и т.д. имеют абстрактную природу: упоминание имени *S* вызывает у слушающего зрительный образ объекта, который и является источником метафоры.

С точки зрения поверхностного синтаксиса, генитивная и инструментальная конструкции представляют собой целое семейство вариантов с разным набором элементов, а именно:

S S'gen *плети рук*	S' Sins *борода лопатой*
A S S'gen *иссохшие плети рук*	S' A Sins *борода узким винтом*
	S' V Sins *борода разлетелась веером*
	S' V A Sins *борода торчит белым веником*

При этом в генитивной конструкции имя характеризующего объекта *S* является хозяином второго имени существительного *S'* (ср. *плети → рук*), в инструментальной конструкции направление зависимости обратное (ср. *лопатой ← борода*). В инструментальной конструкции также может возникать глагол (ср. *пойти, лежать, завиваться, воткнуть, держать* и т. п.), который выступает посредником между двумя именами, то есть управляет как именем *S* (в творительном падеже), так и именем *S'* (в именительном или винительном падеже).

Имя прилагательное не просто распространяет именную группу *S'*: лексемы, которые заполняют эту позицию, принадлежат к ограниченному кругу лексических классов (прежде всего, прилагательные формы, консистенции, цвета и температуры). Кроме того, присутствие глагола и прилагательного часто ослабляет ограничения на сочетаемость имен *S* и *S'* (ср. Десятова и др. 2008; см. также ниже). Вследствие этого разновидности конструкции с распространителями, согласно теории Грамматики Конструкции, претендуют на то, чтобы считаться отдельными конструкциями. Тем не менее, поскольку вопрос о статусе конструкций непринципиален для целей настоящего исследования, мы будем назы-

* Работа выполнена при частичной поддержке РФФИ, грант № 10-06-00586.

[1] Также в роли эталонов форм могут выступать геометрические объекты, ср. *треугольник, линия, круг, крест* и др.

вать генитивной или инструментальной конструкцией любой из вышеперечисленных поверхностно-синтаксических вариантов.

Вопрос, который интересует нас в данной статье, совсем о другом. Зачем в языке для выполнения одной функции (обозначения формы одного объекта через другой объект) имеются конструкции с двумя разными типами падежного управления? Далее в работе мы опишем черты, общие для обеих конструкций, а затем проанализируем их различия, прежде всего, разницу в лексическом заполнении слотов. На этом основании будет сделан вывод о семантических доминантах и расхождениях в профилировании пространственной ситуации.

Инструментом описания служит топологическая классификация русской предметной лексики, разработанная Е. В. Рахилиной (Рахилина 2000) и автором данной статьи.

2. Топологическая классификация лексики

В рамках когнитивной семантики употребление предметных имен в пространственных конструкциях описывается с помощью понятия зрительной, или пространственной схемы (spatial schema, Talmy 1983, 2005) . Талми исходит из предположения, что человек, воспринимающий и описывающий пространство, отвлекается от многих геометрических особенностей видимых объектов и оперирует ограниченным набором гештальтов, таких как «контейнер», «линия», «(плотная) среда» и др. К этим идеализированным типам апеллируют пространственные схемы, закрепленные за теми или иными языковыми средствами. Например, пространственная схема предлога *along* апеллирует к «линии», схема предлога *through* — к «среде» или «проходу» и т. п.

Как инструмент лингвистического описания, пространственная схема должна, с одной стороны, объяснять запреты на сочетаемость, а с другой стороны, предсказывать те или иные тонкости пространственной интерпретации. Первый случай иллюстрирует пример из Талми *The string of beads hung *over / against the wall* . Схема предлога *over* (в одном из употреблений) предполагает соположение двух параллельных плоскостей; если объект нельзя отнести к категории «плоскости», то конструкция с *over* невозможна. Второй случай подробно рассмотрен в работе Рахилина 2000 на примере конструкций с прилагательными формы и размера. В частности, пространственная схема прилагательного *круглый* допускает несколько типов объектов: «трехмерный объект», «пластина», «стержень», «выступ» и т. п. В зависимости от того, как категоризуется предмет, называемый существительным, меняется интерпретация всей конструкции, ср. соответственно *круглый комок* (шар), *круглый пирог* (плоский объект, имеющий форму круга), *круглый столб* (стержень, имеющий форму круга в поперечнике), *круглые щеки* (полу-шар).

В нашем исследовании мы выделяем следующие типы топологических классов (см. также Десятова и др. 2008, Махова и др. 2009):

- поверхности (*спина, живот*; сюда же относятся разнообразные участки земли: *луг, огород* и др.),
- пластины (*ладонь, бумага, листок, записка, тетрадь, блин*),
- выступы (*брови, нос, груди, щеки, рот, губы, борода, ушки, живот*),
- углубления (*яма, траншея*),
- отверстия (*дыра, окно*),
- полосы (*дорога, линия, шов*),
- стержни (*палка, оглобля, палец, свечка*),
- дуги (непрямые стержни: *брови, крылья, мост, радуга*),
- столбы (вертикально ориентированные стержни: *ноги, мачта*),
- веревки (гибкие стержни: *волосы, косички, хвост, хобот, шланг*),
- кольца/круги (*браслет, блин*),
- шары (*голова*, *ядро*),
- оболочки (*рубаха, сарафан, ткань, штаны, сапоги, голенища*),
- верхние части (*стрижка, прическа, шляпа*),
- вместилища (*здание, котел*),
- трехмерные объекты (неопределенной формы: *подушка, диван, книжка, колода, чемодан*).

Диагностиками топологической аффилиации имен служат ограничения на их употребление в разных типах пространственных конструкций: атрибутивной с прилагательными размера и формы (*толстые щеки, круглые щеки*), препозитивной (*поверх барьера*), с глаголами и именами, характеризующими форму (*подбородок торчит*, *губки бантиком, брови в ниточку*) и ряд других. Существенной особенностью топологической классификации является способность имен входить в несколько классов, ср. *нос* как выступ (*острый нос*) и полоса (*прямой нос*): это связано с тем, что форма объекта может оцениваться по-разному в зависимости от способа видения пространственной ситуации.

3. Материал исследования

Исследование строится на данных Национального корпуса русского языка. На первом этапе был обработан список контактных двусловных коллокаций НКРЯ вида «имя существительное + имя существительное в род. пад.» и «имя существительное + имя существительное в твор. пад.», встречающихся чаще 2 раз, из которого было отобрано множество сочетаний предметных имен с подходящей семантикой формы. При этом были исключены из рас-

Таблица 1. Примеры заполнения позиций S и S'.

S		*S'*	
генитивная	инструментальная	генитивная	инструментальная
а) **лопаты** рук	книжка **гармошкой**	кишка **брандспойта**	**удочка** дугой
б) **скала** пансионата	живот **горой**	горб **горы**	**горы** горбами
в) **горошины** глаз	нос **картошкой**	шарики **маслин**	**пироги** сердечком
г) кишки **шлангов**	жгуты **кишками**	гора **живота**	**живот** подушкой
д) **козырек** бровей	ладонь **козырьком**	щели **бойниц**	**каблуки** стаканчиком
е) **конус** трубы	труба **конусом**	нитка **линии**	**линия** горбом
ж)	столб **пальмой**	нитки **берез**	**кипарисы** языками
з) **змейка** дороги	дорога **змейкой**	палки **рыб**	**кот** веером

смотрения коллокации, относящиеся к другим конструкциям с родительным и творительным падежом, например, посессивная (ср. *веревки корней*[2] *— веревка палача*), с отношением «часть — целое» (ср. *крючок носа — крючок коромысла*), квантификации (ср. *лопаты рук — лопата снега*), параметрическая (ср. *веревки корней — веревка шагов* [*в двести*], *красавец лицом*), конструкции сравнения, не использующие метафору формы (ср. *юбка колоколом — половник колоколом* [*гремел в кастрюле*], *сталь небес*) и др. Полученное множество коллокаций было дополнено материалами из РГ 1980. По этим данным был составлен список имен, которые могут выступать в роли имен эталона формы *S* (около 400 единиц).

На втором этапе по корпусу были отобраны и проверены все примеры употреблений указанных имен в генитивной и инструментальной конструкциях формы, в частности, содержащие и такие комбинации имен, которые не попали в исходный список коллокаций. Заметим, что не рассматривались контексты, обозначающие форму части объекта (ср. *крючок коромысла*; эти примеры были отнесены к категории «часть — целое»), а также контексты, описывающие траекторию движущегося объекта и способ движения, ср. *веер пуль, ручейки слез, мяч пулей влетел в ворота*; в том случае, если в примере описывалось изменение формы объекта (*ср. лук согнулся дугой*), критерием отбора служило то обстоятельство, что имя *S* описывает конечную форму всего объекта. Общий объем данных составил порядка 7000 контекстов.

4. Лексико-семантические классы в слотах конструкций

Каждая конструкция характеризуется ограничениями на заполнение ее слотов. Для того чтобы понять сходства и различия генитивной и инструментальной конструкций, проанализируем, какие лексико-семантические группы способны заполнять позиции *S* (имя объекта эталона формы), *S'* (имя характеризуемого объекта), *A* (имя прилагательное, стоящее при имени *S*) и *V* (глагол).

В этом разделе будут описаны свойства, общие для обеих конструкций. В позициях *S* и *S'* могут выступать имена любых онтологических и функциональных классов предметной сферы, например, имена инструментов (а), природных объектов (б), кондитерских изделий и плодов (в), частей тела (г) и прочих частей (д), геометрических фигур (е), растений (ж), животных (з) и т. п. Позицию характеризуемого объекта *S'* могут замещать также имена лиц (ср. *Ольгерд лежал неподвижной* горой), в то время как в позиции *S* преобладают неодушевленные имена, из одушевленных здесь представлены только некоторые имена животных (ср. *бабочка, ежик, змея, уточка*), а имена лиц не встречаются вовсе. Сдвиг спектра можно легко объяснить: чтобы объект служил эталоном формы, его видимые очертания должны быть достаточно простыми, даже примитивными; неподвижность объекта также служит гарантией того, что форма не будет изменяться. Имена одушевленных объектов используются в метафорах сравнения, но совсем другого типа, например, если речь идет о человеческих качествах (*смотреть победителем*), движении (*взмывать птицами*) и т. п.

Более существенные ограничения на заполнение слотов конструкций касаются соответствия топологических классов имен *S* и *S'*. Здесь можно выделить три случая: согласование (уподобление), квантификация и рассогласование.

1. **Согласование.** В Десятова и др. 2008 была отмечена тенденция к совпадению или уподоблению классов в позиции *S* и *S'*. Например, если имя *S'* называет объект дугообразной формы, то его партнером в позиции *S* скорее всего будет имя того же топологического класса, ср. *серп месяца, брови дугами/полумесяцем, дуга радуги, дуга Троицкого моста*. «Выступы» уподобляются «выступам» (ср. *нос горбом, сугроб горой, гоб горы*), «веревки» — «веревкам» (*руки плетьми, невидимая нить веревки*), «стержни» — «стержням» (*камыши ресничками*) и т. д. В более слабом варианте имя в позиции *S* принадлежит к близкому топологическому классу:

[2] Первой приводится коллокация с семантикой формы.

так, близки друг другу удлиненные вытянутые объекты (ср. классы «полосы», «дуги», «стержни», «веревки», «столбы»), круглые объекты (ср. классы «кольца/круги» и «шары»), плоские объекты (ср. классы «поверхности» и «пластины»). Поскольку топологические характеристики источника и мишени метафоры совпадают, механизм сопоставления состоит в том, чтобы усилить воздействие визуального образа. Неслучайно при этом имена эталона формы часто выступают в сопровождении прилагательных «визуального ряда», а именно, прилагательных цвета и света (ср. 1–2) и формы (ср. 3–5):

(37) **Белые нитки корней**, свисающие с потолка, бесплотно касаются лба и ушей [А. Иванов. Географ глобус пропил (2002)];

(38) По **сверкающей нитке шоссе** божьей коровкой полз красный автобус [И. Грекова. На испытаниях (1967)];

(39) **Брови тоненькими дугами**, не иначе выщипывает [В. Мясников. Водка (2000)];

(40) **Тонкая ниточка** его **пробора**, всегда тщательно расчесанного в любое время дня и ночи, когда Звягинцеву приходилось видеть Горохова, сейчас сбилась куда-то в сторону, исчезла в волосах, спутанных на макушке [А. Чаковский. Блокада (1968)];

(41) Высоко вверху, почти на гребне стены, охватывавшей привольную низину, **острым клином** выдавался **выступ** — последняя ступень подъема. [И. А. Ефремов. Час быка (1968–1969)].

Для усиления визуального образа используются также глаголы класса «фиктивного движения» (Talmy 1996), например, глаголы типа *пойти*, *бежать, вилять, спускаться*, *тянуться*, отсылающие к траектории движения взгляда воображаемого наблюдателя, сопровождают имена удлиненных объектов, а глаголы типа *вздыматься, выступать, торчать* — имена выступов, ср. (6–7):

(42) На загривок холма змеиной **лентой вползает шоссе** и убегает по стрелке, на которой написано «Древний Акрополь» [Д. Каралис. Роман с героиней // «Звезда», 2001];

(43) На следующий день мы пошли мимо **гор** Кату, **поднимавшихся** зубчатым **горбом** среди холмов [В. А. Обручев. В дебрях Центральной Азии (1951)].

2. **Квантификация.** С точки зрения топологической классификации, имена веществ и множеств объектов представляют особый нейтральный случай. Эти имена не несут презумпции формы, поэтому в конструкции с родительными и творительным падежом они употребляются с именами эталона формы *S* любого топологического класса (ср. *ниточка дыма, слезы горошинками, гора арбузов, арбузы горой*). Речь идет о квантификации неоформленной субстанции, и имена S берут на себя роль имен кванторов.
3. **Рассогласование**. Если имена *S* и *S'* относятся к разным топологическим классам, то налицо конфликт двух презумпций о форме объекта. Побеждает здесь форма *S*, которую механизм метафоры «навязывает» характеризуемому объекту *S'*. Стандартный случай — модификация формы гибких объектов, например, гибких пластин или полос, ср. (8–9):

(44) Борис сунул руку в карман пальто и, не разжимая своей пухлой и нежной, как его лицо, **ладони**, **протянул ее горстью** в сторону собеседника [В. Громов. Компромат для олигарха (2000)];

(45) Заглядевшись на эти красоты, я потерял бдительность и заблудился в **петлях горной дороги**. [Э. Розенталь. Чудаки с планеты Ко // «Вестник США», 2003.06.25].

Другой интересный случай — аккомодация топологических типов (Десятова и др. 2008). Например, в примере (10) сочетаются классы «выступ» и «круг/кольцо». В принципе, колесо — это изолированный трехмерный объект, но поскольку грудь как выступающая часть тела не может быть представлена как нечто изолированное от поверхности тела, то ее форма приобретает свойства колеса лишь частично (можно сказать, что в профиль такая грудь похожа на часть колеса).

(46) А путанка Клавдия Ивановна, пятипудовая женщина, сидела у самовара, распаренная, в тренировочных штанах, **грудь колесом**, размалеванная и в бигуди [Л. Измайлов. Объезд по кривой (1988)].

В примерах (11–12) имя *S'* (*залив*) относится к классу «вместилищ», тем не менее, залив здесь уподобляется трубке и дуге. Очевидно, что в (11) речь идет о проходе, по которому должны пройти корабли, а в (12) — о части, ограниченной берегом. Таким образом, вся конструкция в целом обозначает форму функционально выделенной части объекта *S'*.

(47) Вокруг него весело перекрикивались его матросы, которым новое дело понравилось, а перед ним с высоты мачты открывалось за **узкой кишкой** Финского **залива** просторное Балтийское море, и в глубине его, возле Киля, уже выстраивались в походный ордер германские линкоры, крейсера, миноносцы. [Л. С. Соболев. Капитальный ремонт (1932–1962)];

(48) На твоих холмах повиснут сады, в кронах чинар, тополей, вязов потонет мозаика твоих крыш, скроются башни Климента и Криско; санатории обегут **дугу залива**, и возникнет здесь шумный, тенистый, до блеска промытый курорт, проходящие стальные фрегаты будут заполнять свои трюмы целыми озерами мягкой украинской воды. [Ю. Черниченко. Небесная глина (1968) // «Юность», 1969].

5. Две конструкции: точки расхождения

В предыдущем разделе мы рассмотрели свойства, присущие обеим конструкциям. Тем не менее, собранный материал показывает, что наборы лексем, заполняющих позиции *S* и *S'*, в генитивной и инструментальной конструкции не совпадают. Иными словами, не всякая комбинация имен, возможная в генитивной конструкции, может быть употреблена в инструментальной конструкции и наоборот. Лексические профили конструкций пересекаются, ср. (13–14), (15–16), но лишь частично.

(49) … маленький разгоряченный мужичок в измятом **пузыре** нейлоновой **рубахи** увлек стряпуху рывком в визг и дребезг пляшущего праздника [О. Славникова. Стрекоза, увеличенная до размеров собаки (1995–1999)];

(50) Перед крыльцом съезжей — спины, от ветра вздутые **пузырями рубахи**, выдубленные солнцем голенища шей, галдеж, гомон [Е. И. Замятин. Рассказ о самом главном (1923)];

(51) Старая Махотиха, Лешкина мать, обморочно всплеснула вялыми **плетьми рук**, закрылась ими и завыла, завыла, терзая всем души, уткнув черное лицо в черные костлявые ладони [Е. Носов. Усвятские шлемоносцы (1977)];

(52) Никита стоял, понурив голову, сдвинув плечи, повесив **плетьми руки** и поставив ступни носками немного внутрь [В. М. Гаршин. Денщик и офицер (1880)].

В общем и целом, инструментальная конструкция характеризуется существенно большим разнообразием лексических комбинаций, чем конструкция с родительным падежом. В частности, это связано с тем, что в первой конструкции часто присутствует глагол, ср. *поле поднималось горбом, океан вздымается огромной водяной горой между Европой и Америкой*. Без глагола комбинация *поле горбом, океан горой* была бы вряд ли возможна: именно глагол вносит идею трансформации формы из состояния A в состояние B.

Однако, генитивная и инструментальная конструкции различаются также и по списку нераспространенных словосочетаний вида *S S'*, ср. *губы трубочкой*, но **трубочка губ*, *галстук бантиком*, *ноги бутылками*, но *?бутылки ног*, но *#бантик галстука*. Чтобы понять природу ограничений на сочетаемость лексических элементов в конструкции, нужно вспомнить, что каждая из рассматриваемых нами языковых единиц происходит из своей семьи конструкций. Инструментальная конструкция представляет собой частный случай конструкции сравнения, поэтому, в принципе, источником сравнения может быть объект самой изысканной (но, что важно, легко опознаваемой) формы. Генитивная конструкция — родственница конструкций со значением квантификации, указания части целого и параметров. При обозначении формы индивидуальных объектов здесь отдается предпочтение именам простых геометрических форм, таким как *линия, полоса, клин, дуга*. В генитивной конструкции также имеется тенденция к обозначению постоянной формы объекта (обратим внимание, что в сочетании *губы трубочкой* трубочка — это форма, которую губы принимают на время). Такие сочетания, как *бантик галстука* будут интерпретированы скорее как «часть — целое», в то время как соответствующее сочетание с творительным падежом *галстук бантиком* будет обозначать разновидность галстука.

Различаются также лексические профили заполнения позиции *A* — прилагательного при имени *S*. В генитивной конструкции наблюдается особый подкласс адъективных распространителей S, обозначающих материал изготовления объекта S', ср. *железная кишка теплотрассы*, *гранитная нитка набережной*. Мы видим здесь уже двойную метафору: свойство быть сделанным из железа или гранита приписывается воображаемому объекту S. Очевидно, эта особенность конструкции происходит из семантико-синтаксической прозрачности, свойственной большой семье генитивных конструкций в целом (ср. конструкцию меры *выпить стакан молока* и др.).

6. Заключение

Итак, мы рассмотрели две разновидности конструкций со значением формы, генитивную и инструментальную. Объектом нашего внимания были синтаксические особенности, а также сходства и различия в лексическом профиле конструкций, то есть в особенностях лексического заполнения конструкционных элементов. Если говорить о генитивной и инструментальной конструкции как о представителях больших семей, то можно увидеть, что сравнение по форме — это лишь одна из немногих точек, где сходятся функции родительного и творительного падежа. Каждая конструкция имеет свой прототип, что неизбежно накладывает отпечаток на способ профилирования ситуации, а значит, набор типов реальных и воображаемых объектов, которые могут попадать в сферу действия метафоры сравнения, оказывается разным.

Литература

1. *Гилярова К. А.* Языковая концептуализация формы физических объектов. (2002). Дисс… канд. филол. наук. М.: МГУ.
2. *Десятова А. В., Ляшевская О. Н., Махова А. А.* (2008). Конструкция с творительным формы «X Y-ом» // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2008». Вып. 7 (14). М.: РГГУ.
3. *Кобозева И. М.* (2000). Как мы описываем пространство, которое видим: форма объектов // Нариньяни А. С. (ред.), Труды международного семинара «Диалог 2000» по компьютерной лингвистике и его приложениям. Т. 1. Протвино. С. 155–161.
4. *Махова А. А., Ляшевская О. Н., Десятова А. В.* (2009). Части тела с точки зрения топологии: корпусное исследование // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2009». Вып. 8 (15). М.: РГГУ.
5. *Рахилина Е. В.* (2000). Когнитивный анализ предметных имен: семантика и сочетаемость. М.: Русские словари.
6. *РГ 1980* — Русская грамматика. (1980). М.: Наука.
7. *Talmy L.* (1987/2000). How language structures space // Talmy L. Toward a cognitive semantics. V. I. Cambridge, MA: MIT Press.
8. *Talmy L.* (1996/2000). Fictive motion in language and ‘ception’ // Talmy L. Toward a cognitive semantics. V. I. Cambridge, MA: MIT Press, 2000.
9. *Talmy L.* (2006). The fundamental system of spatial schemas in language // Hampe B. (ed.), From perception to meaning: Image Schemas in Cognitive Linguistics. Mouton de Gruyter.

Конечно в повседневном общении (по материалам ЗКРЯ «Один речевой день»)

Konechno [of course] in everyday communication (in the speech corpus “One speech day”)

Маркасова Е. В. (markasovaelena@yandex.ru),
Воробьева С. А.(sophie@wau.spb.ru)

Санкт-Петербургский государственный университет, Санкт-Петербург

Слово *конечно* имеет разные функции в повседневном общении. В докладе на материале звукового корпуса «Один речевой день» рассматривается проблема распределения функций этого слова в живой речи, приводятся сведения, отражающие степень индивидуализированности его использования в речи информантов. Уделяется внимание проблеме идентификации коммуникативных задач информанта с помощью перцептивного эксперимента по распознаванию слова конечно, вычлененного из речевого потока.

1. Вводные замечания

Слово *конечно* давно привлекает внимание исследователей (Баранов, Кобозева1984, Баранов, Плунгян 1993, Киселева1994, Шмелева 1995 Шаронов 2009 и др.). В зависимости от аспекта исследования ученые называют его либо вводным словом, либо дискурсивным словом, либо словом-предложением, либо утвердительной частицей, либо коммуникативом. Семантика *конечно* детально описана К. Л. Киселевой и Д. Пайаром (1994), затем К. Л. Киселевой (1998), и это описание, на наш взгляд, не нуждается в уточнениях и не требует дополнений. «Наиболее типичными контекстами появления *конечно* являются монолог (Мне, *конечно*, не следовало говорить ему об этом) и ответ на вопрос (Ты придешь? — *Конечно*!)» (Киселева 1998:345). Независимо от того, имеем ли мы дело с вводным словом (как в первом случае) или со словом-предложением (коммуникативом), оно способно выражать разные оттенки отношения говорящего к имеющимся фактам или высказанным мнениям, причем используется значительно чаще, чем его синонимы *естественно* и *разумеется*. Так, в НКРЯ *конечно* встречается 94 467 раз (13 879 документов), тогда как *естественно* — только 14 152 (документов: 5304), а для *разумеется* отмечено 559 вхождений на 225 документов.

2. Постановка проблемы

Мы рассмотрим функционирование *конечно* в повседневном общении. Эта *цель* предполагает решение следующих задач: во-первых, необходимо установить, как распределяются функции *конечно* в живой речи (как часто оно встречается в качестве вводного слова и в качестве утвердительной частицы). Во-вторых, важно выяснить, какова степень индивидуализированности его использования. ЗКРЯ «Один речевой день» позволяет выявить информантов, склонных и не склонных к включению в речь этого слова, и в первом приближении установить факторы, влияющие на функционирование этого слова в разных коммуникативных ситуациях.

3. Материал и принципы работы с ним

Основным источником материала является Звуковой корпус русского языка повседневного общения «Один речевой день», создаваемый в СПбГУ и частично включенный в НКРЯ (Богданова 2009, Sherstinova 2009). Всего сейчас расшифровано 33,85 часов звучащей речи (125 звуковых файлов, около 225 тысяч слов). При подсчете наиболее часто встречающихся в ОРД слов выяснилось, что слово

конечно зафиксировано 355 раз, то есть встречается не так часто, как *вообще* (689), но чаще, чем *блин* (305), *короче* (197), *наверное* (182), *в общем* (176) (Sherstinova 2009). Если иметь в виду функционирование *конечно* как заменителя *да* или *нет*, есть смысл учитывать, что слово *да* встречается 4370 раз, а слово *нет* — 1508.

Для описания функций слова *конечно* были просмотрены все существующие расшифровки, затем примеры на основе восприятия письменного текста были разделены на три группы: 1) вводное слово, 2) слово-предложение с утвердительным значением, 3) слово с семантикой, не дифференцируемой на уровне читательского восприятия, т.е. без прослушивания. Затем материал анализировался по дикторам: выявлялись дикторы, склонные и не склонные к использованию этого слова в общении. Определяющим показателем здесь было количество *конечно* в ходе общения с собеседниками. Сведения о встречаемости этого слова в речи коммуникантов сопоставлялись с данными, характеризующими речь информантов.

Следующим этапом работы было прослушивание звуковых файлов, содержащих слово *конечно*, и попытка выявления формальных характеристик, способствующих распознаванию разных интенций говорящего слушающим.

4. Функции *конечно* в повседневном общении

При прочтении расшифрованных записей оказалось, что только 178 случаев можно охарактеризовать как вводные слова в «классическом» употреблении: они выражают различные оттенки отношений к происходящим событиям или мнениям собеседников и используются в прямом значении: «естественно, разумеется». Например:

(53) ну вот / *конечно* это создало определённые проблемы // *П ну в том числе (э…э) в общем / *П обстановка / *П криминальная / в Финляндии / (И42)

(54) о / а если этот цвет ? (Р1)
ну / можно *конечно* / да // (И37)

47 случаев можно считать коммуникативами в значении «да». Причем интересно отметить, что однозначное отнесение определенных реплик к этому типу было возможно лишь по отношению к тем информантам, с которыми автор статьи знаком лично или речь которых сам расшифровывал. В этих случаях знание индивидуальной манеры говорения дает основания с уверенностью говорить о том, какой смысл вкладывал информант в это слово. Например:

(55) здесь цвета как бы побольше … # *конечно* // (Р1 # И37)

(56) коне… / это у всех бывает / а почему нет ? *конечно* ! # я их перемерила // в нескольких магазинах // (И24)

107 употреблений не могут быть описаны на основе прочтения расшифровки, несмотря на широкий контекст и понятность ситуации. Они могут служить как для выражения иронического согласия (*конечно* = «еще чего»), так и для выражения иронии или сарказма по отношению к чужой уверенности по какому-либо поводу. Иногда невозможно сказать, является ли *конечно* синонимом *да*, или же вводным словом в незаконченном предложении. Кроме того, они могут служить для имитации чужой речи или передачи собственной речи в прошлом. Эта имитация трудна для анализа, поскольку возникает наложение разновременных интенций говорящего. С одной стороны, это интенция, отнесенная в прошлое и связанная с другим контекстом, а с другой стороны, — современная, включающая в себя информацию о прошлом и оценку говорящего, его поведения или события, ориентированную на настоящий контекст и нового слушателя. Это обстоятельство заставило нас отдельно рассматривать такие случаи.

Без слухового анализа и привлечения к прослушиванию сторонних наблюдателей решить, каковы функции включения в речь подобных *конечно*, просто невозможно. Например:

(57) нет / я даже не рассматривала эту свечку // он что-то ткнул / я сказала / да-да *конечно* (И04)

Без прослушивания звукового файла *конечно* может иметь разные интерпретации: «естественно» — плеонастическое (из-за «да-да») интенсивное выражение согласия, ироничное осуждение («еще что придумал!»).

Сложно анализировать *конечно,* используемое в качестве знака солидаризации с коммуникантом: распознавание этой функции требует обращения к ситуации.

Еще сложнее говорить о семантике *конечно* в тех случаях, когда в левом контексте есть *да, ну* или *нет*.

(58) слушайте / да *конечно* ! (И09-коммуникант).

(59) ну *конечно* // да *конечно* / а чего ? ну я думаю / что мы так это / посидим уж // (коммуникант — И13)

(60) Да нет *конечно* (И19)

В ходе обработки материала мы пытались с помощью перцептивного эксперимента решить вопрос

о том, распознаются ли слушателем коммуникативные интенции говорящего на основании контекста, или следует считать, что слово *конечно* несет смысловую нагрузку, сообщающую тексту определенный фон: служит показателем иронии в речи говорящего, обладает специфическими характеристиками, позволяющими распознавать прямое значение (=да; = естественно, разумеется).

В эксперименте приняли участие 10 человек, ни один из которых не имел отношения к филологии и не был связан с ОРД. Для информантов была разработана анкета, в которой типы значений сознательно смешивались с общими коммуникативными характеристиками. Это было сделано для того, чтобы никто из информантов не имел возможности логическим путем установить результат, в котором может быть заинтересован исследователь.

Для прослушивания были отобраны следующие примеры:

(61) да / *конечно* (И11) 112 ms

(62) *конечно* (И11) 56 ms

(63) ну *конечно* (И11) 157 ms

(64) только *конечно* (И11) 68 ms

(65) *конечно* хорошее пиво (И11) 127 ms

(66) форевер *конечно* 128 ms

Участникам опроса были предложены варианты характеристики слова *конечно:*

1. *естественно*, *разумеется* (в буквальном смысле)
2. *безусловно!* (эмоциональное утверждение)
3. ироничное «*естественно*, *разумеется*»
4. синоним слова «*да*», выражение согласия
5. «*да*» ироничное, в значении «*еще чего*», «*что еще придумаешь*»
6. нейтральная передача чужой речи
7. имитация речи человека, который чем-то неприятен
8. другое

Суждения участников опроса были разнообразны (один из них прокомментировал свои ответы: «Вопрос в соотношении степени уверенности и контекста. Чем больше контекст, тем больше уверенность!»). Оказалось, что все отметили последний случай как вызывающий негативные эмоции («с претензией», «имитация речи человека, который неприятен», «ироничное», «выпендривается»). В рамках нашего исследования это означает, что можно говорить о возможности выявления коммуникативных намерений говорящего без учета широкого контекста — лишь по фонетическим характеристикам реализованного слова.

5. Степень индивидуализированности включения в речь *конечно*

Анализ данных ЗКРЯ показал, что слово *конечно* по-разному функционирует в повседневном общении информантов. Чтобы были очевидны различия информантов в этой области, мы решили разбить сводную таблицу, отражающую особенности включения в речь *конечно,* на отдельные таблицы. Кроме пола, длительности в минутах и количества словоупотреблений1, в таблицах отражено общее количество *конечно*, количество слов в минуту, характеризующее интенсивность общения, доля *конечно* относительно общего количества слов, а также — отдельно — количество *конечно* в речи информанта и коммуникантов.

Сгруппировав информантов, близких по времени проанализированной речи, мы получили возможность более четко увидеть соотношение количества словоупотреблений в звуковых файлах одинаковой или примерно одинаковой протяженности. Благодаря этому становится понятно, что, во-первых, общее количество словоупотреблений в речевом отрезке и количество *конечно* и — величины, не связанные отношениями зависимости. Казавшееся аксиоматичным представление о том, что эти показатели будут прямо пропорциональны, оказалось ложным. Например, у информантов И02 и И15 (Табл. 1) на равное количество минут при почти равном количестве словоупотреблений общее количество *конечно* одинаково (2), а у И44 количество словоупотреблений почти в 2 раза выше, но при этом количество *конечно* по-прежнему 2. В файлах И18 и И43 «много» *конечно* (10 и 7 соответственно), однако это ничего не значит: в речи информантов никакого роста или снижения количества *конечно* нет. ОРД пока не дает возможности дифференцировать статистику словоупотреблений информанта и коммуникантов, поэтому введение в наши таблицы данных об употреблении *конечно* на тысячу слов мы посчитали неоправданным.

Есть группа информантов, которые избегают слова *конечно* в повседневном общении: например, И29 обходится без этого слова на протяжении 31 минуты, И01 — на протяжении 93 минут. Далее идут информанты И07 — 97 минут, И08 — 64 минуты, И09 — 31 минута расшифровок. Эти информанты использовали слово *конечно* по одному разу.

В *табл. 1* разброс количества слов в минуту (от 68,95 до 215,39) не сопровождается аналогичными колебаниями в количестве *конечно* у информанта.

---

[1] Нами использованы сведения из доклада: А. С. Асиновский, Н. В. Богданова, Е. В. Маркасова, А. И. Рыко, С. Б. Степанова, Т. Ю. Шерстинова «Звуковой корпус русского языка: заключительная стадия формирования», сделанного в рамках семинара на факультете филологии и искусств СПбГУ 20.12.2009.

Таблица 1. *Конечно* в повседневном общении информантов (23–24 мин.)

Информант	Пол	Длительность в минутах	Кол-во словоупотр.	Кол-во слов в минуту	Доля *конечно*	Количество *конечно*		
						Всего	В речи информанта	В речи коммуникантов
И02	м	23	2234	91,13	0,089	2	1	1
И15	м	23	2440	106,08	0,081	2	2	0
И18	ж	23	3172	137,91	0,317	10	2	8
И43	м	24	1655	68,95	0,422	7	1	6
И44	м	23	4954	215,39	0,040	2	0	2

Таблица 2. *Конечно* в повседневном общении информантов (29–31 мин.)

Информант	Пол	Длительность в минутах	Кол-во словоупотр.	Кол-во слов в минуту	Доля *конечно*	Количество *конечно*		
						Всего	В речи информанта	В речи коммуникантов
И09	ж	31	4865	156,93	0,143	7	1	6
И17	м	30	2827	94, 23	0,212	6	3	3
И20	ж	31	4254	137, 22	0,188	8	1	7
И25	м	29	4303	148,37	0,185	8	6	2
И29	м	31	2594	83,67	0	0	0	0
И41	ж	29	3050	105,17	0,032	1	1	0

Таблица 3. *Конечно* в повседневном общении информантов (36–38 мин.)

Информант	Пол	Длительность в минутах	Кол-во словоупотр.	Кол-во слов в минуту	Доля *конечно*	Количество *конечно*		
						Всего	В речи информанта	В речи коммуникантов
И03	ж	38	5675	149,34	0,052	3	3	0
И10	м	36	3993	110,91	0,150	6	0	6
И36	м	41	3858	94, 09	0,311	12	8	4

Таблица 4. *Конечно* в повседневном общении информантов (55–59 минут)

Информант	Пол	Длительность в минутах	Кол-во словоупотр.	Кол-во слов в минуту	Доля *конечно*	Количество *конечно*		
						Всего	В речи информанта	В речи коммуникантов
И06	ж	57	5242	91,96	0,095	5	3	2
И12	ж	57	6526	114,49	0,122	8	3	5
И13	ж	55	8126	147,74	0,319	26	20	6
И21	м	56	5938	106,03	0,168	10	8	2
И37	ж	59	4070	68,98	0,467	19	13	6

Таблица 5. *Конечно* в повседневном общении информантов (62–67 мин.)

Информант	Пол	Длительность в минутах	Кол-во словоупотр.	Кол-во слов в минуту	Доля *конечно*	Количество *конечно*		
						Всего	В речи информанта	В речи коммуникантов
И08	ж	64	7290	113,9	0,068	5	1	4
И11	ж	65	6636	102,09	0,165	11	8	3
И22	ж	64	7905	123,51	0,063	5	2	0
И27	ж	66	7767	117,68	0,102	8	2	6
И28	м	67	7592	113,31	0,197	15	7	8
И30	ж	62	6821	110,02	0,161	11	8	3

Таблица 6. *Конечно* в повседневном общении информантов (73–78 мин.)

Информант	Пол	Длительность в минутах	Кол-во словоупотр.	Кол-во слов в минуту	Доля *конечно*	Количество *конечно*		
						Всего	В речи информанта	В речи коммуникантов
И19	ж	75	8858	116,9	0,079	7	3	4
И35	м	73	12671	173,48	0,142	18	4	14
И42	м	78	7322	93,87	0,205	15	12	3

Таблица 7. *Конечно* в повседневном общении информантов (93–97 мин.)

Информант	Пол	Длительность в минутах	Кол-во словоупотр.	Кол-во слов в минуту	Доля *конечно*	Количество *конечно*		
						Всего	В речи информанта	В речи коммуникантов
И01	ж	93	11206	120,49	0,0357	4	1	3
И07	м	97	10962	113,01	0,0547	6	1	5

В *таблице 2* отражены колебания в интенсивности общения (от 83,67 до 156,93 слов в минуту), причем интересны И25 (род занятий неизвестен), являющийся доминирующим участником диалога с девушкой: он 6 раз произносит *конечно*, описывая свое путешествие в Хельсинки.

Среди информантов в *таблице 3* выделяется И36 (менеджер, работает бригадиром монтажников), при сравнительно низкой интенсивности общения он довольно часто произносит *конечно*.

Группа информантов *таблице 4* наиболее интересна для анализа: в таблице отражена разница в количестве *конечно*, произнесенных информантами и коммуникантами. Так говорят И13 (20 из 26 за 56 минут), И21(8 из 10 за 56 минут), И37 (13 из 19 случаев за 59 минут).

В *таблице 5*: И30 — студентка или аспирантка университета, в речи *конечно* — синоним *естественно*. И11 — выпускница филологического факультета, преподает, функции *конечно* разнообразны, часто используется при имитации чужой речи.

В группе, представленной в *таблице 6,* выделяется И42, кандидат наук, филолог, в настоящее время работает консультантом по проектам. В его речи слово *конечно* в основном заменяет *да* в домашних разговорах.

В группе (*таблица 7*) отметим функции *конечно* в речи коммуникантов: в файлах с записями И07 это слово постоянно произносят женщины, от которых зависит продвижение документов И07: *конечно* становится показателем речевого доминирования. В записях И01 *конечно* говорит зубной врач, то есть тоже есть ситуация коммуникативного неравноправия.

Таблица 8. *Конечно* в повседневном общении информантов (120–125 мин.)

Информант	Пол	Длительность в минутах	Кол-во словоупотр.	Кол-во слов в минуту	Доля *конечно*	Количество *конечно*		
						Всего	В речи информанта	В речи коммуникантов
И04	ж	125	20 892	164,3	0,153	32	8	24
И24	ж	120	9405	78,77	0,127	12	11	1

В *таблице 8* представлены данные И04 (разговоры родственников и друзей в день рождения): здесь *конечно* выглядит как элемент стиля общения в микроколлективе, очень часто выступает показателем не просто согласия, но и солидаризации собеседников. И24 производит впечатление человека, критически настроенного и пытающегося «приобщить» собеседника к своей системе оценок, чему способствует часто используемое *конечно*.

Выявляется интересная закономерность: очень редко можно говорить о «равноправии» информанта и коммуникантов в повседневном общении: если *конечно* часто произносит информант, то в речи коммуниканта оно встречается реже, и, соответственно, наоборот. Неясными в этом плане остаются файлы И04, И19, И28, И29. Это редкие случаи условного «равноправия» информантов и коммуникантов. Пока не удалось выяснить, какими словами информанты, обходящиеся без слова *конечно*, заменяют его в живой речи.

Вопреки мнению о распространенности *конечно* в монологах, материалы ОРД показывают, это слово постоянно встречается в диалогах, а среди информантов выделяется группа тех, кто очень часто включает это слово в свою речь с разными интенциями. Это И11 (8 из 11 за 55 минут), И42 (12 из 15 случаев за 78 минут), И24 (11 из 12 за 120 минут), И04 (8 из 36 за 120 минут), И37. Всех этих информантов объединяет наличие коммуникантов, являющихся родственниками, подругами, приятелями. Кроме того, слишком частое (на общем фоне) включение в речь *конечно* оказывается чертой «домашних» полилогов, средством интимизации общения, показателем солидаризации с собеседником, знаком речевого доминирования, что можно установить на основе описанных в ЗКРЯ эпизодов.

Многие информанты, принадлежащие к этой группе, имеют прямое отношение к работе с людьми: это либо преподавательская деятельность, либо сфера торговли, менеджмента: И11 — молодая девушка, выпускница филологического факультета, преподаватель, менеджер, И13 — молодая девушка, сотрудница фирмы, связанной с архитектурой, И21 — молодой человек (до 29 лет), менеджер, И37 — женщина, художник, репетитор, И42 — мужчина (около 50 лет), по специальности филолог, И24 — женщина (около 50 лет), преподаватель, историк, И04 — молодая женщина, преподаватель, филолог. Опираясь на эти сведения, мы сможем вернуться к анализу функций *конечно* в речи каждого из отмеченных информантов, чтобы выяснить: какие ограничения на функции *конечно* накладывает характер отношений коммуникантов и информантов, как один и тот же информант произносит это слово в условиях нормальной (равноправной) коммуникации и при асимметричном распределении коммуникативных ролей. Мы предполагаем, что высокая частотность слова *конечно* обусловлена родом деятельности информанта, но это не главный фактор: большее значение имеет коммуникативный статус говорящего и иерархия статусов других участников коммуникации.

Литература

1. *Баранов А., Кобозева И.* Вводные слова в семантической структуре предложения //Системный анализ значимых единиц русского языка: Синтаксические структуры. Красноярск. 1984.
2. *Баранов А., Плунгян В., Рахилина Е.* Путеводитель по дискурсивным словам русского языка. Москва. 1993.
3. *Богданова Н. В., Асиновский А. С., Русакова М. В., Рыко А. И., Степанова С. Б., Шерстинова Т. Ю.* Звуковой корпус как способ мониторинга и фиксации разных форм естественного языка // Компьютерная лингвистика и интеллектуальные технологии. Вып. 8(15) По материалам международной конференции «Диалог 2009». С. 38–45.
4. *Дискурсивные* слова русского языка: опыт контекстно-семантического описания (1998) К. Киселева — Д. Пайар. Москва. Наука.
5. *Киселева К., Пайар Д., Разлогова Е.* КОНЕЧНО, или «Чужая правда» // Русистика сегодня. 1994. №1.
6. *Шаронов И. А.* Коммуникативы и методы их описания // Компьютерная лингвистика и интеллектуальные технологии. Вып. 8(15) По материалам международной конференции «Диалог 2009». С.543–548.
7. *Шмелева Т. В. Диалогичность модуса // Вестник МГУ, Сер. 9. Филология. 1995. №5.*
8. *Sherstinova T.* The Structure of the ORD Speech Corpus of Russian Everyday Communication. // *«Text, Speech and Dialogue» TSD-2009*. LNAI 5729. Springer-Verlag Berlin Heidelberg. 2009. pp. 258–265.

Непрочитанный доклад

Александра Семёновича Нариньяни

Голубые города

26 апреля 2010 года ушел из жизни Александр Семенович Нариньяни. Его смерть всколыхнула воспоминания о более чем полувековом знакомстве, профессиональном сотрудничестве и дружбе с этим неординарным, замечательным ученым и человеком.

Так случилось, что в 1955 году мы оказались с Александром Семеновичем (далее с Сашей) в одном доме и в одном подъезде, ездили в одном и том же лифте. Тогда никто из нас, видимо, не подозревал, что нам — казалось, весьма далеким по будущей профессии — предстоит вступить в весьма близкий научный контакт. Я учился на филологическом факультете МГУ на отделении классической филологии, а он — в МИФИ, сугубо техническом вузе. Однако каждый из нас проделал довольно существенный, непрямолинейный путь к занятиям, связанным с той областью, которая ныне называется компьютерная лингвистика.

Саша провел свою дипломную практику в новосибирском Академгородке, а после окончания МИФИ в 1962 году резко изменил место своего проживания и отправился создавать и обживать Академгородок, суливший новые горизонты для академической науки. Там он начинает в 1962 году свою академическую карьеру в должности старшего лаборанта в Институте математики, после открытия ВЦ переходит туда.

Академгородок славился своим вольнодумством и активной культурной жизнью. Там действовал клуб «Под интегралом», где выступали те артисты и барды, которым был закрыт доступ к зрителям и поклонникам в столицах, организовывались небывалые художественные выставки, проводились диспуты в условиях небывалой демократии. И там кипела активная научная жизнь. Возможно, под влиянием своего научного руководителя академика Андрея Петровича Ершова, усиленно занимавшегося информатикой и информатизацией, Саша обратился к лингвистике как науке, связанной с информационными и интеллектуальными технологиями. Это потребовало от него довольно существенного погружения в лингвистику.

Я вновь встретился с ним в 70-е годы, но теперь — как с вполне владеющим лингвистической терминологией, проблематикой и менталитетом. Это было время инициированного им (теперь уже более тридцати лет назад) проекта междисциплинарного семинара, объединявшего лингвистов, инженеров, специалистов по искусственному интеллекту, психологов и даже философов. Этот уникальный ежегодный зимний семинар проходил в 1975–1989 гг. в Эстонии, особенно при участии коллег из Тартуского университета. Постепенно эти собрания приобретали черты тематических семинаров — «Модели общения», «Диалог». Последнее название сохранилось и по сей день.

Саша был генератором глобальных проблем и успешным продюсером больших проектов. В течение более десяти лет он сотрудничал с нашей университетской группой в форме тогдашних грантов, по тем временам — хоздоговоров. Не имея никаких собственных финансовых ресурсов, он ухитрялся доставать их у тогдашних меценатов — различных министерств, которым советский Госплан выделял фонды на науку. Для меня вся эта его деятельность была полной загадкой, мы имели дело только с ним и отчитывались только перед ним. Формат хоздоговоров позволял создавать ставки и расширять кадровый состав группы.

В 70–80-е годы мы очень сблизились с Сашей. Мы много спорили на наших workshop'ах и мозговых — как он любил говорить — штурмах, но наши споры не были никогда омрачены отстаиванием личных амбиций. Главной целью был добрый поиск правильных решений.

90-е годы были для Саши трудным испытанием. Еще в 1989 году он оставил Академгородок, мало походивший на тот, с которого все начиналось. Я знаю об этом понаслышке и понимаю только общую траекторию изменений. Саша вернулся в Москву, увлеченный идеей создания Голубого города, где были бы воплощены усохшие идеалы Академгородка. Он чуть ли не подыскал в Подмосковье место для этого Города и сколачивал команду единомышленников. Но стране было уже не до Голубых городов. Когда рухнула старая система, все привычные рычаги влияния исчезли. Саша создал в 1999 году ведомственный Институт искусственного интеллекта, способ существования которого был мне непонятен, но в новую ситуацию вписаться было трудно. Он мыслил широкомасштабными проектами на десятилетия, а искать и находить средства к сиюминутному проживанию не было сильной стороной его интеллекта. Мало кто знает, что последние 10 лет, продолжая работать, он фактически не получал зарплаты.

Все мы сейчас еще не готовы осмыслить всё, что было сделано Сашей. Но мне кажется, что главная Сашина черта — романтизм: вера во всесилие добра и доброго знания, в возможность делать благое дело безо всякой корысти и пересчета на возможные личные дивиденты. Он, по существу, жил в неком виртуальном мире, где все устроено просто — заботиться не о себе, а об идее. В этом была его сила и его слабость.

В последние годы чувствовалось, что он угнетен многими несбывшимися ожиданиями. Однако его постоянно пытливая мысль, направленная на идеальную цель с далеко идущими последствиями продолжала напряженную работу. Публикуемая в материалах конференции его последняя статья оказалась как бы его завещанием. В этой статье утверждается идея государственного масштаба. Нынешнее состояние науки пока еще дает возможность России вытянуть счастливый билет и сделать прорывный скачок в наиболее продвинутой у нас в стране области знания. Именно этот билет есть естественная и доступная для России сфера получения довольно быстрого роста в экономике и культуре, требующая инвестиций не в нефтепроводы и даже не в нанотехнологии, а в человеческие мозги и научные знания.

Саша говорит власти: «Используйте реальный шанс, если вас заботит будущее России как процветающего современного государства. А я ничего не могу сделать больше, чем указать на эту возможность, моя миссия выполнена, я могу уходить».

А. Е. Кибрик
3 мая 2010

Русский язык как национальная программа

Нариньяни А. С.

1. Язык как фундамент культуры

Язык — необходимый компонент базового треугольника любой мировой культуры, нет прецедентов, когда бы язык разрушался, а большая цивилизация оставалась.

Однако меняются этапы развития человечества и формирующие их технологии-доминанты. Сегодня будущее любой культуры определяется уровнем ее симбиоза с информационно-коммуникативными технологиями (ИКТ), и прежде всего погружением ее языка в глобальное Интернет-пространство, из которого на наших глазах формируется электронная «нервная система» современного человечества.

Если до сих пор та или иная страна определялось своей географической территорией и ролью в истории, то теперь борьба перешла во всеобщую информационную сеть, где развертывается следующий этап соревнования за место в глобальном мироустройстве.

Сегодня обработка информации на естественном языке является основой практически любого вида деятельности и поэтому ее роль особенно важна для подавляющего большинства приложений, тем более что в ближайшей перспективе широкое внедрение текстовых и речевых технологий сделает взаимодействие с компьютером по-настоящему массовым и естественным.

Для России компьютерные технологии обработки языковой информации (КТОЯ) важны еще и для сохранения быстро сокращающейся виртуальной русскоязычной территории, определяющей ее связи с многомиллионной диаспорой за рубежом, а также для поддержки тех международных коммуникационных контактов, которые сложились за вторую половину двадцатого века.

КТОЯ уже пришли в государственную и общественную сферы, становятся частью дома и личной жизни. Таким образом, обработка языковой информации и по важности и по масштабу – проблема национального уровня. Однако именно ее масштабность ставит под вопрос ее посильность для основного большинства языков и стран. Лишним доказательством этого является тот факт, что пока только небольшое число языков успешны в некоторых секторах приложений. Причем даже эти результаты можно оценить как достаточно умеренные, несмотря на огромный объем вложенных здесь усилий и средств.

2. Национальная программа

2.1. В известной степени обсуждаемая ситуация сходна с положением в атомной физике на границе 30х и 40х годов, когда перспектива появления атомного оружия потребовала крайней мобилизации ресурсов — прежде всего, интеллектуальных — для форсированного перехода от *исследовательской науки* к дорогой и сложной области оборонной *индустрии*.

Можно упрекнуть меня в чрезмерной драматизации. Однако сегодня, как и тогда, решается ключевой вопрос судьбы страны, определяется будущее нашей культуры в ряду основных мировых культур. И сегодня снова необходим рывок, обеспечивающий переход от науки, привыкшей работать в режиме НИР, в национальную отрасль, способную не только масштабно решать прикладные задачи, но активно работать на экспорт.

В таком прорыве мы можем опираться: на отечественные школы лингвистики и информатики, которые остаются пока одними из лучших в мире. Реализация этого потенциала должна позволить создать те инновационные языковые технологии, которые обеспечат необходимый успех в борьбе за ведущее место России в глобальном распределении ролей.

2.2. Очевидно, что координация усилий такого масштаба возможна только на уровне Национальной Программы, обеспечивающей самую продуктивную кооперацию участвующих в ней научных, прикладных и коммерческих сил.

Такая Программа (обозначим ее «Русский Язык-ИКТ») должна совместить четыре ключевые составляющие:

- ⇨ *фундаментальные работы по интеграции лингвистических ресурсов русского языка;*
- ⇨ *комплекс прорывных программных КТОЯ, в первую очередь, для русского языка;*
- ⇨ *механизмы маркетинга создаваемых продуктов на отечественном и мировом рынке;*
- ⇨ *подготовку высококвалифицированных кадров и обучение пользователей.*

В стране сегодня работает несколько программ поддержки работ в области русского языка, однако они ограничены финансово, не координированы и ориентированы на достаточно специальные сектора. Эти работы могут служить дополнением к предлагаемой Программе или быть ее частью, но никак не ее альтернативой.

Обеспечить ядро Программы может только консорциум фирм и научных коллективов, способный при участии инвесторов и поддержке государства обеспечить ее реализацию — создание комплекса интеллектуальных ИКТ нового поколения, ориентированных на эффективную обработку информации на русском и, возможно, ряде других языков.

2.3. Значение Программы для общего фронта работ, определяющих сохранение нашей страной быстро утрачиваемого ею статуса научной сверхдержавы, переоценить нельзя, поскольку эффективность КТОЯ и базирующихся на них современных каналов обмена информацией является не менее важной составляющей прогресса, чем уровень кадров и современное оборудование.

Актуальность и стратегическая важность проблемы подтверждается широким фронтом работ в данной области и объемом средств, которые тратятся на них во всех ведущих странах. Однако, текущий уровень основной массы таких продуктов не обеспечивает пока результатов, гарантирующих их прикладную эффективность и их массовое внедрение.

Потенциальные участники такой Программы имеют многолетний опыт в этой области, подтверждающий как высокий уровень разрабатываемых ими технологий, так и оригинальность положенных в их основу ноу-хау. Именно это сочетание опыта, принципиально новых подходов и сложившихся высококвалифицированных команд и, главное, имеющихся заделов, обеспечивают Программе высокую вероятность прорыва в создании качественной технологии следующего поколения.

В судьбе страны Программа «Русский Язык-ИКТ» должна выполнить не менее ключевую функцию, чем проекты, включаемые руководством страны в спектр стратегических инновационных направлений. Если в развитии нано- и био-технологий Россия может сотрудничать со всем миром, то место русского языка в ИКТ — это будущее *нашей страны* и ее культуры — определяется только здесь совокупностью именно наших усилий.

Потеря времени будет усложнять эту задачу на порядки, а каждый выигранный год даст прикладной и моральный выигрыш, намного превышающий объем инвестиций, требуемый для этого необходимого рывка сегодня.

3. Структура Программы

3.1. Структура проблемной области КТОЯ в самом общем виде может быть представлена следующим образом:

1. *Содержание и понимание (смысла текста, знания об области приложения, онтологии)*
2. *Анализ текста*
   - 2.1. Индексация и классификация текстов
   - 2.2. Извлечение содержания текста
   - 2.3. Понимание сообщения конкретной тематики
   - 2.4. Понимание текста пользователя в контексте диалога
   - 2.5. Определение синтаксиса и стиля и их корректировка
   - 2.6. OCR
3. *Генерация текста*
   - 3.1. Вопрос, Сообщение, Команда, Ответ в контексте диалога
   - 3.2. Документ конкретной тематики и жанра
4. *Перевод (автоматизированный и автоматический)*
5. *Поисковые системы*
   - 5.1. Файловые системы
   - 5.2. Однородные текстовые массивы, БД, электронные архивы и библиотеки
   - 5.3. Плохо структурированные массивы
   - 5.4. Локальные сети
   - 5.5. Интернет
6. *Диалог «человек — компьютер»*
7. *Лексика*
   - 7.1. Словари одноязычные, двуязычные, многоязычные

7.2. Корпусные БД
7.3. Тезаурус
8. *Гипертексты*
9. *Стандарты*

3.2. Естественно, что такая общая классификация не может претендовать на полноту, поскольку должна покрывать весь спектр проблематики от теории до коммерческих продуктов.

Фундамент Программы образуют три основных взаимосвязанных группы работ, отражающие следующие направления тематики КТОЯ:

– Формальные описания основных составляющих знаний о языке,
– Систематизированная лексика (словари),
– Аппарат представления знаний и его приложения в КТОЯ.

Каждая из этих групп рассмотрена ниже в соответствующем подразделе.

3.3. Формализация лингвистических знаний о языке — необходимый компонент создания КТОЯ, основные составляющие лингвистического обеспечения которого включаю такие разделы как:

- *Морфология*
- *Лексическая семантика*
- *Синтаксис*
- *Структура текста*
- *Структура документа*
- *Организация диалога.*

Каждая из этих составляющих складывается из:

⇨ разработки формальных средств описания,
⇨ самого описания соответствующей части лингвистических знаний,
⇨ программных средств использования этих знаний в КТОЯ.

Для первой половины этих — наиболее проработанных и широко используемых — составляющих сегодня особо важными являются лингвистические НИОКР, направленные на доработку полного их описания, поскольку именно оно определяет достаточность и эффективность соответствующих компонентов КТОЯ. При этом обеспечение соответствующих формальных и программных средств может рассматриваться как задача относительно техническая. Для второй половины списка дополнительных усилий требует разработка всех трех компонентов.

3.4. Лексика — совокупность слов и устойчивых словосочетаний того или иного языка, определенной сферы деятельности или конкретной совокупности текстов, такая же ключевая составляющая КТОЯ, как и остальные две. Основные компоненты данного раздела обеспечивают несколько уровней работ от накопления языкового материала (первые два раздела) до формирования словарей самого верхнего уровня (тезаурусы), организующих лексику по семантическому принципу с отражением базовых отношений. Данные компоненты включают:

- *Корпус русского языка со всеми его более специальными составляющими,*
- *Проблемная лексика (каталоги, справочники, терминологические словари и т. п.),*
- *Словари одноязычные, включая базовый интегральный словарь русского языка,*
- *Словари двуязычные / многоязычные,*
- *Тезаурусы.*

Эта подгруппа проектов включает как содержательное наполнение перечисленных составляющих лексического фонда КТОЯ, так и разработку обслуживающих их технологий. Таким образом, как и в предыдущей подгруппе, каждая из этих составляющих складывается из:

⇨ разработки формальных методов для соответствующих групп лексики,
⇨ содержательного наполнения,
⇨ программных средств оформления лексики в компоненты КТОЯ.

Масштаб работы и распределение весов указанных трех частей для перечисленных составляющих различаются не менее чем на порядок и зависят от:

- степени их проработанности,
- полноты и объема содержательного наполнения,
- необходимости каждой из них для тех или иных приложений.

Практически все перечисленные компоненты являются необходимыми и для речевых технологий, особенно для ограниченных предметных областей, где необходимо учитывать предметную направленность текста при разрешении неоднозначности распознавания.

3.5. Наиболее эффективные КТОЯ связаны с использованием смысла ЕЯ сообщения, который в свою очередь неотъемлем от контекста этого сообщения, т. е. знаний о *прагматике дискурса* и о *предметной области*, к которым оно имеет отношение. Поэтому третьей необходимой составляющей КТОЯ является как сам аппарат представления знаний, так и его применение в соответствующих областях. Сюда относятся:

- *Технология представления знаний,*
- *Онтологии,*
- *Модели предметной области,*
- *Представление смысла текста.*

Естественно, что технологии извлечения смысла текстового сообщения включают все эти компоненты, которые должны войти как необходимые направления в состав Программы.

4. Базовые составляющие Программы

4.1. Таким образом, комплекс прорывных КТОЯ должен охватить все основные составляющие языковых

ИКТ следующего поколения, включающие такие ключевые области приложений как:

- *Содержательная обработка потоков текстовых сообщений,*
- *Эффективный интеллектуальный поиск,*
- *Качественный документооборот,*
- *Обработка текстовых данных в системах поддержки принятия решений,*
- *Электронные архивы и библиотеки нового поколения,*
- *Значительное повышение качества машинного перевода,*
- *Интеллектуальные системы в образовании и медицине,*
- *Комплекс речевых технологий,*
- *Диалоговые языковые интерфейсы к прикладным системам,*
- *Понимание текстов в ограниченной предметной области,*
- *Системы двойного назначения и др.*

Понятно, что для успеха Программы она должна опираться на такие базовые составляющие как:

→ Компьютерный словарный фонд русского языка,
→ Качественные технологии анализа и синтеза текстовой и речевой информации,
→ Технологии понимания текста.

Ниже две составляющие будут рассмотрены более детально.

4.2. *Компьютерный словарный фонд русского языка.* Здесь в основу Программы входят три базовых проекта:

• «Корпус» — работы по созданию корпуса русского языка ведутся коллективом энтузиастов (текущий объём более 140 млн. слов) при минимальном финансировании. Программа требует доведения сделанного до готовности по объёму и качеству, а также по возможности дальнейшего развития и пополнения.

• «Словарь» — интегральный словарь русского языка. Сегодня электронных словарей русского языка разного качества многие десятки, но все они далеки от полноты и завершенности. Для приложений и для самой лингвистики необходим единый комплексный «государственный» словарь, включающий как современную лексику с детально проработанной стандартизованной морфологией, так и все имеющиеся словари в качестве дополнительных специализированных приложений. Система такого масштаба требует концентрации соответствующих усилий для ее создания, развития и пополнения, в частности, и для перехода на следующих этапах к формированию многоязычной версии.

• «Тезаурус русского языка» по типу известных тезаурусов других языков. Значение подобного издания для нашей культуры, лингвистики и КТОЯ невозможно переоценить. Стоит упомянуть издательство Оксфордского университета, подготовившего к публикации «Исторический тезаурус Оксфордского словаря английского языка», самый большой подобный словарь в мире: 800 тысяч значений 600 тысяч слов, организованные в 354 категории и 230 тысяч подкатегорий. Идеографический словарь О. С. Баранова, представляющий собой практически единственный на сегодняшний день большой тезаурус русского языка, может послужить основой такого национального издания. Представляется, что при наличии средств за два — три года можно подготовить первый выпуск, причём сразу на трёх уровнях: полный, массовый и школьный.

4.3. *Речевые технологии:* Программа должна предусматривать исследования и разработки перспективных методов распознавания, синтеза, сжатия и идентификации русской речи.

Эти задачи включают разработку новой эффективной технологии полного транскрибирования речевого сигнала, в результате которой будут получены алгоритмы для систем обработки речи с характеристиками существенно лучшими и намного более точными, чем существующие в настоящее время. Эта часть Программы ориентируется прежде всего на естественную речь на русском языке и его диалектах по широкому спектру каналов речевой связи, в том числе введенную с микрофона, радиопередач, телефонных разговоров и т. п.

К основным областям, нуждающимся в разных формах речевых технологий относятся:

- Текстовые редакторы (ввод текста голосом),
- Сфера обслуживания (транспорт, торговля, call центры, др.),
- Управление техническими и бытовыми приборами, в частности, телевизорами и мобильными телефонами,
- Обработка речевой информации (извлечение данных, перевод, др.),
- Интерфейс «конечный пользователь — компьютер»,
- Документооборот (речевые материалы, тексты дискуссий, др.),
- Образование, прежде всего — обучение языкам,
- Каналы связи, в частности, мобильная связь, и многое другое.

5. Общая архитектура Программы

Очевидно, что целью Программы является не только интересующие разработчиков компоненты теории и технологии проблем обработки текста на естественном языке. Ее стратегическая цель — решение задач рынка, т. е. фронта соответствующих приложений.

Именно это определяет общую архитектуру Программы, которая, так или иначе, сводится к уровням

систем, покрывающим основные направления этого фронта. Эти уровни можно обозначить как:

- Базовые технологические составляющие КТОЯ, которые были кратко рассмотрены в п.4,
- Программные подсистемы, используемые как крупные строительные блоки систем обработки текста,
- Системы категории 1,
- Макросистемы,

В следующих разделах приводятся примеры компонентов этих уровней.

6. Программные компоненты систем КТОЯ

6.1. К уровню программных подсистем, используемых как крупные строительные блоки систем обработки текста, можно отнести компоненты КТОЯ, выделенные в следующие классы:

Корректоры текста. Технологии автоматического или автоматизированного исправления текста, включающие, в зависимости от выполняемых функций и их сложности, операции разного уровня, к которым относится:

- Коррекция орфографии,
- Коррекция пунктуации и синтаксиса,
- Коррекция стиля.

Автоматическая либо автоматизированная коррекция разного качества уже используется достаточно широко и может быть частью большинства приложений.

6.2. *Средства «предобработки».* Функции этой категории являются важными составляющими многих, если не большинства, приложений КТОЯ; они включают:

- Идентификацию языка входного текста,
- Нормализацию (унификацию) текста,
- Интеллектуальный OCR.

Идентификация языка сообщения может использоваться в речевых технологиях, возможно, наряду с предварительной классификацией речевого сообщения (определение жанра или ориентации ПО) как средство выбора соответствующего проблемного словаря

6.3. *Лингвистические процессоры* (анализаторы текста, парсеры). Программные модули, выполняющие основные функции КТОЯ, связанные, как правило, со следующими составляющими процесса обработки текста:

- Морфологический анализ,
- Синтаксический анализ,
- Анализ структуры текста.

В общем случае все эти компоненты лингвистических процессоров представляют собой тандемы <исполнитель, система правил>.

6.4. *Генерация (синтез) текста.* Перевод в текст формального представления информации. Таким представлением может быть:

- конструкция из типовых составляющих,
- синтаксическая структура текста (в основном, в МП),
- содержание, выраженное формальными средствами представления смысла (семантики и прагматики) генерируемого текста,
- сочетание этих уровней представления.

Приложениями, использующими генерацию текста в качестве подсистемы, являются:

- → машинный перевод,
- → диалоговые системы (генерация реплик компьютера),
- → подсистемы генерации сообщений Баз знаний, экспертных и диагностических систем, систем поддержки решений и др.,
- → подсистемы ответа на запрос для хранилищ данных (оформление ответа в виде ЕЯ-текста или комментарии к выдаваемым данным).

Кроме перечисленных приложений, имеется перспективный сектор КТОЯ, в котором основной функцией является именно генерация:

- → Типовых писем в деловой переписке,
- → Типовых документов в широком спектре от коротких готовых текстов, требующих подстановки нескольких параметров, до текстов большого размера (отчеты, типовые юридические документы, технические инструкции и др.)

Генерация текста должна использоваться во всех приложениях, связанных с использованием синтеза речи.

6.5. *Семантические процессоры* (извлечение содержания и понимание текста). В связи с повышением сложности и разнородности текстов растет необходимость в интеллектуализации технологий их обработки. К категории Семантических процессоров здесь отнесены КТОЯ, обеспечивающие в разной степени анализ и извлечение содержания обрабатываемого текста. В этом плане можно выделить четыре уровня интеллектуализации, связанные с учетом содержания документа, а именно:

- *Классификация текстов, формирование онтологий,*
- *Содержательная индексация текстов,*
- *Извлечение компонентов содержания,*
- *Понимание текста.*

Перечисленные типы процессов, так же как и лингвистические, представляют собой тандемы <исполнитель, система правил>.

- *Классификация текстов:* технология определения принадлежности текста к одной или нескольким категориям из заданного (возможно, расширяемого) набора категорий \ классов; как правило, строится на базе статистических методов. Формирование Онтологий является

наиболее продвинутой технологией создания содержательной структуры корпуса текстов, имеющей ключевое значение для остальных трех уровней интеллектуализации.

- *Содержательная индексация текстов:* технология обработки текста, основной функцией которой является повышение эффективности поиска в массиве текстов с помощью создания индекса, — структуры, сопоставляющей объектам поиска (текстовым компонентам) выделенные в них ключевые слова и словосочетания, маркирующие те объекты, в которых эти ключевые элементы присутствуют. Таким образом, ключевой элемент становится виртуальным адресом того подмножества текстовых компонентов, которые ему сопоставлены. В частности, индекс может служить базисом классификатора текстов.
- *Извлечение компонентов содержания* — один из основных элементов интеллектуализации технологий обработки текстов, роль которого быстро растет в связи с увеличением объемов и усложнением структуры информационного пространства, в частности, за счет роста сложности и разнородности самих документов. В зависимости от сложности задач, относимых к автоматическому или автоматизированному извлечению компонентов содержания из компьютеризированной информации (в основном, из текста, реже из БД и других форм данных) под извлечением содержания могут пониматься:
  - ⇨ Выделение значимых компонентов текста (даты, референты и т. п.),
  - ⇨ Выявление упоминаний о фактах и событиях из конкретного набора,
  - ⇨ Реферирование (выделение основного содержания),

Трудность создания эффективных технологий этого ряда возрастает с повышением сложности извлекаемого содержания. Более простые технологии используются практически во всех интеллектуальных системах обработки текста.

- *Понимание текста* — анализ содержания ЕЯ-текста и перевод его на формальный язык представления смысла в проекции на контекст анализируемого сообщения, т. е. *прагматику дискурса* и знания о *предметной области*, к которым оно имеет отношение. *Полное понимание* пока достигается только в экспериментальных системах и при анализе ЕЯ-запроса для *узких* (ограниченных) *предметных областей*.

Все четыре составляющие этой группы технологий являются необходимыми для развития речевых технологий, поскольку качество анализа речи непосредственно связано с выявлением ее содержания.

6.6. *Организация диалога «человек — компьютер».* Обмен репликами (команды, вопросы, сообщения) между пользователем и компьютером, направленный на решение конкретной проблемы: постановка задачи, уточнение ее условий, оценка результатов и т. п. С расширением сферы использования компьютеров и спектра решаемых задач диалог становится необходимой составляющей всей системы информатизации.

Сегодня возможности диалога ограничиваются в основном технологией меню, которая успешно обслуживает четко структурированные и ограниченные классы команд и запросов, но перестает быть эффективной при усложнении диалога и росте числа альтернатив.

Возможность *диалога на естественном языке* пока остается перспективой, близость которой для различных приложений определяется сложностью задачи и предметной области. ЕЯ-диалог требует уровня развития, на который соответствующие технологии выходят только в самое последнее время.

Освоение внеязыковых средств только начинается, но в будущем даст возможность значительно расширить возможности диалога, в частности, для учета эмоциональной составляющей пользователя (в частности, контроля состояния оператора в критических ситуациях), обучения иностранному языку, доступа к компьютеру слабослышащих или глухонемых людей, и т. п.

Области приложений диалога:

- Запрос к информационным системам,
- Интерфейсы к прикладным системам.

С развитием речевых технологий сфера устной речи сможет не только полностью покрыть сектор применения ЕЯ диалога (текст останется необходимым только там, где речь неприменима по техническим причинам), но и расширить его за счет тех приложений, где руки оператора заняты и\или набор текста непродуктивен .

7. ПРИЛОЖЕНИЯ

7.1. Приложения КТОЯ можно достаточно условно разделить по сложности на системы категории 1 и макропрограммы. К примерам систем первой категории отнесем следующие классы:

- Хранилища текстовых данных,
- Документооборот,
- Электронный архив,
- Электронная библиотека,
- Текстовый редактор.

Системы этой категории ориентированы на создание технологий, являющихся как самостоятельными продуктами, так и ключевыми компонентами макропрограмм.

7.2. Раздел макропрограмм объединяет наиболее крупные системы, ориентированные на создание стратегических продуктов и технологии КТОЯ:

- Корпоративные информационные системы,
- Интеллектуальные порталы,
- Системы машинного перевода,
- Интеллектуальные поисковые машины,
- Электронная торговля нового поколения и др.

Все макропрограммы, хотя и в разной степени, потенциально связаны с использованием речевых технологий.

8. Рынок технологий обработки текста и речи

В большинстве КТОЯ нуждаются практически все области деятельности, поскольку каждая из них требует таких систем как:

- Качественный документооборот,
- Тематические электронные архивы и информационные справочные системы, в частности, БД технической документации,
- Автоматическая содержательная обработка потоков текстовых сообщений,
- Аналитическая обработка текстовых данных в системах поддержки принятия решений,

Стоит привести несколько примеров областей широкого применения КТОЯ.

☐ Системы корпоративного управления любого уровня от госструктур и крупных корпораций до муниципалитетов и компаний среднего бизнеса нуждаются в массовом использовании КТОЯ, поскольку всем им требуется обработка текстовых документов, связанных со спецификой их деятельности.

☐ Системы образования (школы, ВУЗы, курсы и т. п.) включают те же компоненты, но специализированные в соответствии с особенностями данных секторов приложений. При этом для технологий этого направления важно развитие систем дистанционного обучения.

☐ Вся современная медицина базируется на специализированном документообороте, включающем:

⇨ Ведение историй болезни,

⇨ Выписку справок и рецептов,

⇨ Те же электронные архивы и информационные справочные системы, связанные с медицинскими справочниками самого разного типа,

⇨ Системы обучения, повышения квалификации и т. п.

Развитые медицинские учреждения нуждаются в экспертных и диагностических системах, разработка которых требует наличия баз знаний и аналитической обработки текстовых данных в системах поддержки принятия решений. Фактически поликлиника и больница должны превратиться в текстовый конвейер, автоматизированный во всех узлах контакта с человеком (пациент, врач, администратор).

☐ Системы двойного назначения для силовых министерств нуждаются в системах корпоративного управления, системах обучения и подготовки кадров, а также в таких технологиях как:

⇨ Автоматическая содержательная обработка потоков текстовых сообщений,

⇨ Эффективный интеллектуальный поиск,

⇨ Машинный перевод,

⇨ Комплекс речевых технологий,

⇨ Совершенствование связи,

⇨ Криминалистическая экспертиза и т. п.

Заключение

Оптимальной формой реализации Национальной Программы такого масштаба является формирование одного (или нескольких по направлениям) Консорциума на основе кооперации коммерческих фирм, научных коллективов, вузов и государственных структур, направленного на реализацию широкой национальной программы соответствующего профиля.

За последние десять лет автор участвовал в предварительной работе по подготовке нескольких аналогичных программ, которые, к сожалению, по разным причинам до реализации не доходили.

Соответствующие материалы, значительно более детальные по проработке, были использованы при подготовке этой статьи, в задачи которой входило дать представление о полноте и масштабе Программы, а не отразить ее с соответствующей полнотой и детализацией.

Хотелось бы надеяться, что время для реализации Программы пришло, поскольку без нее вряд ли можно планировать всерьёз такие национальные процессы как модернизация и инновация.

Кореферентные отношения в тексте — сравнительный анализ размеченных данных

Coreferential relationships in text — comparative analysis of annotated data

Недолужко А. Ю. (nedoluzko@ufal.mff.cuni.cz)

Карлов университет, Прага, Чехия

Данная работа опирается на находящийся в разработке проект разметки именной кореференции и ассоциативной анафоры на материале синтаксически аннотированного корпуса чешских текстов PDT 2.0. В ходе работы над разметкой кореференции выяснилось, что относительно низкое соответствие между разметчиками объясняется не столько их ошибками и невнимательностью, сколько объективной неоднозначностью интерпретации текста. В докладе приводится классификация типов и возможных причин несоответветствий и рассматриваются некоторые типовые примеры множественности интерпретаций кореферентных отношений в тексте.

1. Введение

Одним из наиболее актуальных направлений в области обработки естественного языка является в настоящее время извлечение информации из текста. Выявление кореферентных и анафорических отношений из связного текста — необходимый механизм для решения этой задачи.

Проект разметки именной кореференции и ассоциативной анафоры на относительно обширном корпусном материале чешских текстов PDT 2.0 (Prague Dependency Treebank) — один из нескольких десятков мировых корпусных исследований анафорических отношений (напр. Hirschman 1998, Poesio 2004a для английского языка, Recasens — Martí 2010 для испанского и английского, Poesio 2004б для итальянского, Krasavina — Chiarcos 2007 для английского и немецкого и др.). Разметка именной кореференции и ассоциативной анафоры на PDT 2.0 производится вручную и частично автоматически на глубинно-синтаксическом уровне синтаксически аннотированного корпуса чешских публицистических текстов (подробнее с проектом синтаксической разметки можно ознакомиться в Hajičová 2006, Недолужко 2008). В настоящее время размечено около половины всего корпуса PDT 2.0. Регулярно проводятся измерения соответствия между разметками разных аннотаторов (разметчиков) и составляется типология несоответствий, что позволяет делать первые выводы о возможных причинах этих несоответствий. Классификации несоответствий между разметчиками и попытке объяснения некоторых причин их возникновения посвящена данная работа.

2. Типология отношений кореференции и ассоциативной анафоры, размечаемых на PDT 2.0

В PDT 2.0 представлена разметка трех типов кореферентных и анафорических отношений[2]:

1. грамматическая кореференция, где антецедент высчитывается на основе грамматических правил языка. К грамматической кореференции относятся, например, кореференция возвратных местоимений[3], относительных местоимений (*человек, который пьет*), актантов в реципрокальных конструкциях и т. д. Грамматическая кореференция практически никогда не переходит границ предложения

[1] Эта работа была поддержана грантами GACR 405/09/0729 и GAUK 4383/2009.

[2] К более подробному описанию типов см напр. (Nedoluzhko 2007, Nedoluzhko-Mirovsky-Pajas 2009)

[3] В чешском языке возвратное местоимение «ся» всегда является отдельной лексемой (клитикой).

Таблица 1. PDT 2.0 — статистические данные

количество предложений в PDT 2.0	115 844
количество документов в PDT 2.0	7110
размечено грамматической кореференции	100,00 %
размечено прономинальной текстовой кореференции	100,00 %
размечено именной текстовой кореференции	50,00 %
размечено ассоциативной анафоры	50,00 %
количество узлов, связанных кореференцией и ассоциативной анафорой	67 071
процент узлов, связанных кореференцией и ассоциативной анафорой	20,45 %

2. т. наз. текстовая кореференция, где соотношение между кореферентными членами реализуется не только за счет грамматических средств языка, но и на основании знания контекста. Текстовая кореференция может легко переходить границы предложения. Различается прономинальная и именная текстовая кореференция[4].

- прономинальная текстовая кореференция размечается в случае, если в качестве второго члена кореферентного отношения выступают личные и притяжательные местоимения третьего лица, указательное местоимение *этот* в субстантивной функции или эллиптированное и восстановленное на глубинно-синтаксическом уровне местоимение 3-го лица[5]. На глубинно-синтаксическом уровне в PDT эти местоимения восстанавливаются, и им присваивается тектограмматическая лемма #PersPron.
- разметка именной текстовой кореференции является продолжением предшествующей ей разметки прономинальной кореференции. В качестве второго члена кореферентного отношения выступают в основном имена существительные и некоторые наречия *(там, тогда* и др.*)*. В некоторых случаях в отношении кореферентности могут участвовать прилагательные (притяжательные прилагательные, прилагательные, образованные от имен собственных и др.) и числительные (выступающие в субстантивной функции и релевантные для связности текста). При разметке именной текстовой кореференции используется два типа отношений — отношение между конкретнореферентными ИГ (дефолтный тип 0) и отношения между нереферентными и родовыми ИГ (тип NR), причем тип отношения определяется по второму члену анафорического отношения[6].

3. ассоциативная анафора (bridging anaphora), где анафорический член и антецедент уже не кореферентны, но между ними имеется семантическое отношение определенного типа. В настоящее время размечается шесть типов отношений: часть — целое (напр. *Бавария — Германия*), множество — подмножество/элемент множества (*студенты — три студента*), отношение дискурсивного контраста (*Люди не жуют, жуют только коровы*), отношение объекта и его функции/позиции (напр. *школа — учитель*), эсплицитное анафорическое отношение между некореферентными членами (*учителя — такие же учителя*) и отношение «остальное». Последний тип размечается у отношений типа место — житель (*Москва — москвич*), автор — творение, вещь — хозяин, у родственных отношений (*дед — внук*), а также у некоторых предикатно-аргументных отношений (*предпринимательство — предприниматель, спор — участник конфликта* и др.).

Грамматическая и прономинальная текстовая кореференция обработаны полностью на всем корпусе PDT 2.0[7]и в данном докладе учитываются только в статистических данных. Именная кореференция и ассоциативная анафора размечены на половине этого корпуса и являются предметом анализа в данной работе. Некоторые статистические данные о величине корпуса PDT 2.0 и количестве размеченных на нем кореферентных и анафорических отношений представлены в таблице 1.

---

[4] В случае прономинальной текстовой кореференции речь идет, как правило, и об анафорическом отношении. Для именной кореференции размечается именно соответствие референтов данного отношения без учета того, является ли это отношение также анафорическим.

[5] Являясь языком pro-drop, чешский язык имеет сильную тенденцию опускать личные местоимения в анафорических конструкциях (напр. чеш. 0 Nechtěl to říkat. vs. рус. Он не хотел этого говорить.)

[6] Различие ИГ на конкретнореферентные и родовые является свойством независимым от участия данной ИГ в кореферентном отношении. Однако, не имея технической возможности приписывать данный признак всем именным группам, мы ограничиваемся разметкой этого различия только у ИГ, вступающих в кореферентные отношения.

[7] См. Kučová L. и др. 2003

Таблица 2. Соотношение типов именной текстовой кореференции и ассоциативной анафоры

отношение		**количество узлов**	**%**
грамматическая кореференция		11 327	17,54
текстовая прономинальная кореференция		10 747	16,64
текстовая именная кореференция	тип 0	12 034	18,63
	тип NR	2740	4,24
	всего	14 774	22,87
ассоциативная анафора	множество — подмножество/элемент	3307	5,12
	часть — целое	1408	2,18
	дискурсивный контраст	655	1,01
	объект — функция/позиция	355	0,55
	некореферентная анафора	24*	0,04
	остальное	733	1,13
	всего	6482	10,04

* Малое количество отношение типа ANAF объясняется тем, что данный тип стал размечаться только на последнем этапе аннотирования. В ближайшем будущем планируется вернуться к той части корпуса, где этот тип размечен не был, и пройти его еще раз.

Соотношение типов именной текстовой кореференции и ассоциативной анафоры представлено в таблице 2.

3. Измерение соответствий между разметчиками

Аннотирование именной кореференции и ассоциативной анафоры проводится двумя разметчиками с лингвистическим образованием, причем на данном этапе разметка производится «в один слой», т. е. разметчики работают на разных текстах. Тем не менее мы регулярно проверяем соответствие между разметчиками на небольших порциях текстов. В таблице 3 рассмотрено шесть измерений соответствий разметок у разметчиков А и Б, проведенных с приблизительно двухмесячным интервалом. Для измерения соответствия при выборе антецедента как для текстовой кореференции, так и для ассоциативной анафоры, мы использовали F1-measure (Chinchor 1992). Соответствие типов отношений на совпавших парах посчитано в процентах.

Как видно из таблицы 3, успешность разметки не имеет тенденции постоянно возрастать, как бы мы того ожидали. Более того, два первые соответствия имеют по многим параметрам большую успешность, чем последующие измерения 3–5. Успешность последнего шестого измерения вновь несколько возрастает. Тем не менее F1-measure для текстовой кореференции ни в одном измерении не превышает 75,2 %, т. е. для применения наших данных при автоматическом обучении, тестировании и оценке автоматической разметки новых данных их качества еще недостаточно. Еще меньшим является соответствие между разметчиками при установлении отношений ассоциативной анафоры (F1-measure не более 55,5 %). Однако интересно заметить, что при совпадении узлов первого и второго членов отношения кореференции или ассоциативной анафоры (т. е. при наличии одинаковой

Таблица 3. Соответствие разметок у разметчиков А и Б

	1-е изм.	**2-е изм.**	**3-е изм.**	**4-е изм.**	**5-е изм.**	**6-е изм.**
кол-во текстов	3	1	1	2	3	8
кол-во предложений	41	40	101	106	100	211
TKR*, А=Б при выборе антецедента	77,2	63,3	65,6	68,6	62	75,2
TKR, А=Б при выборе антецедента и типа отношения	65,9	39,6	54,5	64,7	50	64,4
TKR, только типы	85,2	62,5	83	94,3	80,5	85,6
bridging**, А=Б при выборе антецедента	55,5	31	35,4	42,2	40,8	43,5
bridging, А=Б при выборе антецедента и типа отношения	55,5	31	33,9	39,1	30	40,9
bridging, только типы	100	100	94,1	92,8	71,4	96,1

* TKR = именная текстовая кореференция.

** Bridging = ассоциативная анафора.

стрелки) совпадение между разметчиками на типе отношений уже достаточно велико (в среднем более 90 %), что свидетельствует о том, что низкая степень совпадений **не** обуславливается слишком сложной типологией размечаемых отношений.

4. К вопросу надежности измерения соответствий между разметчиками

Данные таблицы 3 не представляют стандартной статистической ценности, так как соответствие между разметчиками измерялось на разном количестве предложений различного уровня сложности (см. строки «кол-во текстов» и «кол-во предложений» в таблице 3). Тем не менее они позволяют сделать несколько теоретических наблюдений.

Сравнение параллельных разметок показывает, что **степень соответствия между разметками в высшей степени зависит от величины и сложности текста**. Чем короче текст, тем яснее отношения между его членами, тем выше соответствие между разметками разных аннотаторов. Чем больше в тексте абстрактных понятий, именных групп с родовым денотативным статусом, тем больше вероятность различной интерпретации кореферентных отношений и тем соответствие между разметками ниже. Так для первого измерения использовались три текста длиной не более 15 предложений (см. данные в таблице 3), содержащих в основном только конкретнореферентные ИГ, в результате чего соответствие между разметчиками оказалось достаточно высоким (F1-measure=77,2 % при выборе антецедента и 65,9 % с учетом соответствия при выборе типа на совпавшем отношении). Для второго измерения использовался только один текст, который был однако существенно длиннее предыдущих (40 предложений) и который содержал большое количество родовых и абстрактных понятий. В результате соответствие при выборе антецедента для текстовой кореференции упало до F1-measure=63,3 %, а совпадение на ассоциативной анафоре оказалось совсем низким (F1-measure=31 %). Наиболее надежным является последнее шестое измерение, где для оценки степени соответствия между разметками разных аннотаторов использовалось 8 текстов (211 предложений) различной длины и степени сложности.

5. Типология несоответствий при разметке кореференции и ассоциативной анафоры

Рассмотрим более подробно соответствие между разметками аннотаторов А и Б при последнем (шестом) измерении, причем сосредоточимся только на определении элементов отношений кореференции и ассоциативной анафоры без учета их дальнейшей типологии. Получаем несоответствия трех следующих типов:

- Разметчик А отметил отношение кореференции/ассоциативной анафоры там, где разметчик Б его не увидел, см (1) — в 69 % случаев.

(67) чеш. *Na této stránce vám budeme představovat jednotlivé obory národního hospodářství.* […] *Bylo to v době, kdy se nebývale zvýšil zájem zahraničních turistů a podnikatelů o návštěvu České republiky.*
рус. *На этом сайте будут представлены отдельные отрасли национальной экономики.* […] *Это было в тот период, когда небывало возрос интерес иностранных туристов и предпринимателей к посещению Чешской Республики.*

текстовая кореференция между *národní (национальной)* и *České republiky (Чешской Республики)*:

- разметчик А: отметил отношение кореференции,
- разметчик Б: не отметил это отношение.
- Различный выбор первого или второго члена отношения, см (2) и рис. 1 — в 20 % случаев;

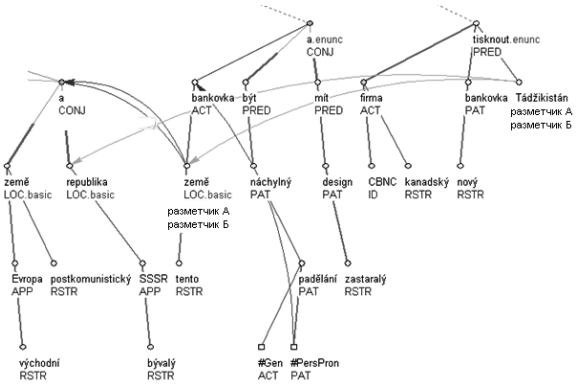


Рис. 1. Различный выбор первого члена отношения

(68) чеш. *Tiskárny bankovek mají i nové zákazníky, především v postkomunistických zemích východní Evropy a republikách bývalého SSSR. Bankovky v těchto zemích jsou náchylné na padělání a mají zastaralý design. Kanadská firma CBNC bude tisknout nové bankovky pro Tádžikistán*
рус. *У монетных дворов есть и другие клиенты, прежде всего в посткоммунистических государствах Восточной Европы и в республиках бывшего СССР. Банкноты в этих странах легко подделать, и у них устаревший дизайн. Канадская фирма CBNC будет печатать новые банкноты для Таджикистана.*

ассоциативная анафора типа «множество — подмножество»:

- разметчик А: отметил отношение «*Таджикистан*» на «*республиках бывшего СССР*»,
- разметчик Б: отметил отношение «*Таджикистан*» на «*в этих странах*», т. е. на всю сочинительную конструкцию «*в посткоммунистических государствах Восточной Европы и в республиках бывшего СССР*».

- Разметчик А отметил отношение кореференции там, где разметчик Б отметил отношение ассоциативной анафоры, см (3) — в 11 % случаев.

(69) чеш. *I přes klesající inflaci ve světě, a tedy nižší potřebu peněz v oběhu, je tisk bankovek a výroba bankovkového papíru jedním z nejlukrativnějších odvětví.* […] … *Rozšíření bankovních automatů vyžaduje neustálý přísun nepoškozených bankovek*.
рус. *Несмотря на снижение инфляции в мире, и соответственно меньшую потребность в оборотных денежных средствах, печать банкнот и производство специальной бумаги является одной из наиболее доходных отраслей.* […] … *В связи с расширением сети банкоматов требуется постоянное пополнение неповрежденных банкнот*.

отношение между *bankovek (банкнот) — nepoškozených bankovek (неповрежденных банкнот)*

- разметчик А: отметил отношение текстовой кореференции типа NR;
- разметчик Б: отметил отношение ассоциативной анафоры типа «множество — подмножество»

6. Причины расхождений разметок разных аннотаторов

Анализ причин расхождений разметчиков при аннотации кореференции и ассоциативной анафоры показал очень интересные результаты. Наиболее вероятной причиной половины расхождений является неоднозначность интерпретации текста. Так например в (3) нельзя точно сказать, кто из разметчиков более прав — тот, кто отметил отношение между *bankovek (банкнот) — nepoškozených bankovek (неповрежденных банкнот)* как ассоциативную анафору типа «множество-подмножество» (т. е. неповрежденные банкноты подмножество всех изготавливаемых банкнот) или тот, кто обозначил отношение между этими именными группами как кореферентное (ведь в сущности все изготавливаемые и выпускаемые в оборот банкноты являются неповрежденными). Подобная ситуация имеет место в примере (2).

Еще в 23 % разметчики расходятся в глубине интерпретации анафорических отношений в тексте, т. е. один из разметчиков отмечает отношение там, где второй считает его уже слишком расплывчатым, что по сути тоже можно рассматривать как неоднозначность интерпретации. Так в (4) отношение между *cestovní ruch* (*туризм*), *hotelových kapacit* (*мест в гостиницах*) и *zahraničních turistů a podnikatelů* (*иностранных туристов и предпринимателей*) может быть интерпретировано как ассоциативная анафора (разметчик А) или как отношение более расплывчатого свойства, которое находится уже за границами разметки (разметчик Б).

(70) чеш. *Méně výnosný cestovní ruch. Hotelových kapacit je mnohem víc než současná poptávka*. […] *Bylo to v době, kdy se nebývale zvýšil zájem zahraničních turistů a podnikatelů o návštěvu České republiky, především Prahy*.
рус. *Менее доходным является туризм. Количество мест в гостиницах существенно превышает современный спрос*. […] *Это было в тот период, когда небывало возрос интерес иностранных туристов и предпринимателей к посещению Чешской Республики*.

Примерно четверть несоответствий может быть интерпретировано как ошибка разметчика (см пример (1)).

Отдельные несоответствия (4 %) являются следствием неточно сформулированных правил разметки.

Таблица 4 обобщает рассмотренные данные.

Таблица 4. Причины расхождений разметок разных аннотаторов

неоднозначность интерпретации	69,00 %
глубина интерпретации	23,00 %
ошибка разметчика	23,00 %
неточность правил разметки	4,00 %

7. Выводы

Большое количество несоответствий между разметками кореференции и ассоциативной анафоры у разных аннотаторов, вызванных неоднозначной интерпретацией данных отношений в тексте, естественным образом приводит к сомнениям в целесообразности осуществления такой разметки на большом корпусе текстов. Маловероятно (хотя мы этого не исключаем и продолжаем над этим работать), что разметка такого уровня сложности может быть в ближайшем будущем использована для успешного тестирования автоматически размеченных данных. Возможно, что исключение из разметки кореференции родовых понятий и девербативов сделает ее более точной, и соответствие между разметками

разных аннотаторов увеличится. Однако это не так просто сделать, так как границы между этими группами весьма размыты[8]. С другой стороны, есть все основания считать такую разметку кореферентной и анафорической информации очень ценной для лингвистических исследований. Анализ размеченного корпуса и сравнение параллельных разметок разных разметчиков дает возможность увидеть некоторые закономерности текста, незаметные при работе одного исследователя над небольшим объемом текстов. Множественность интерпретаций (даже очень простого) текста открывает и помогает решить также много вопросов психолингвистических исследований.

[8] См. Nedoluzhko 2007.

Литература

1. *Cohen J.* A coefficient of agreement for nominal scales. // Educational and Psychological Measurement, 20(1), 1960. С. 37–46.
2. *Chinchor N.* MUC-4 Evaluation Metrics // Proc. of the Fourth Message Understanding Conference, 1992. С. 22–29.
3. *Hajičová E.* и др. PDT 2.0 — Guide. UFAL & CKL, 2006. Доступно на http://ufal.mff.cuni.cz/pdt2.0/
4. *Hirschman L.* MUC-7 coreference task definition. Version 3.0. 1997.
5. *Krasavina O., Chiarcos Ch.* PoCoS — Potsdam Coreference Scheme. Proc. of ACL 2007, Prague, Czech Republic 2007.
6. *Kučová L.* и др. Anotování koreference v Pražském závislostním korpusu. ÚFAL/CKL Technical Report TR-2003-19. 2003.
7. *Nědolužko A.* Zpráva k anotování rozšířené textové koreference a bridging vztahů v Pražském závislostním korpusu. Technical report. Institute of formal and applied linguistics, Charles University, Prague. 2007. Доступно на http://ufal.mff.cuni.cz/~nedoluzko/koref_anot/manual_RK_kratky.pdf
8. *Nědolužko A., Mírovský J., Pajas P.* The Coding Scheme for Annotating Extended Nominal Coreference and Bridging Anaphora in the Prague Dependency Treebank. // Proceedings of ACL-IJCNLP 2009, Linguistic Annotation Workshop (LAW III). Suntec, Singapore, 2009.
9. *Poesio M.* The MATE/GNOME Scheme for Anaphoric Annotation, Revisited. // Proceedings of SIGDIAL. Boston, 2004a.
10. *Poesio M., Delmonte R., Bristot A., Chiran L., Tonelli S.* The VENEX corpus of anaphora and deixis in spoken and written Italian. 2004б. Доступно на http://cswww.essex.ac.uk/staff/poesio/publications/VENEX04.pdf
11. *Recasens M., Antònia Martí M.* AnCora-CO: Coreferentially annotated corpora for Spanish and Catalan. Language Resources and Evaluation. 2010.
12. *Недолужко А., Гаич Я.* Синтаксически аннотированный корпус чешского языка. // Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной международной конференции «Диалог». Выпуск 7 (14) 2008 С. 400–406.

Тайные знаки пунктуации

Secret punctuation marks

Окатьев В. В. (oka@dictum.ru), **Ерехинская Т. Н.** (te@dictum.ru), **Ратанова Т. Е.** (rt@dictum.ru)

ООО «Диктум», г. Нижний Новгород

Вот два петуха,
Которые будят того пастуха,
Который бранится с коровницей строгою,
Которая доит корову безрогую,
Лягнувшую старого пса без хвоста,
Который за шиворот треплет кота,
Который пугает и ловит синицу,
Которая часто ворует пшеницу,
Которая в темном чулане хранится
В доме,
Который построил Джек,,,,,,,,,.

С. Маршак

В работе предложена математическая модель пунктуации, справедливая для различных языков и применимая в различных подходах к синтаксическому анализу. Предложен метод использования разработанной модели в синтаксическом анализе на примере алгоритма Эйснера.

Введение

Существует обширный класс текстов, содержащих грамматически правильные предложения. Это и художественная литература, и публикации в СМИ, и нормативно-правовые документы, и многое другое. Для таких текстов применимы методы формального синтаксического анализа. В настоящее время точность этих методов оставляет желать лучшего. Например, для чешского языка процент правильно построенных деревьев составляет 30,6–35,9 % [11]. Для других языков оценки отличаются ненамного.

Известно, что количество правильно построенных деревьев с увеличением длины предложения уменьшается, как указано в [19], до 9 %. Известно также, чем длиннее предложение, тем больше в нем знаков препинания [7]. Очевидна важность учета пунктуации в синтаксическом анализе [1, 9]. Однако, знаки препинания в компьютерном анализе похожи на занозы — мелкие, но очень неудобные.

Обзор подходов к анализу пунктуации приведен нами в [8]. Дополнительно отметим, что в статистических методах анализа [11, 13, 16, 19] знаки препинания могут лишь влиять на веса связей при машинном обучении. Подходы на основе грамматик [17, 18] позволяют описывать пунктуацию достаточно подробно, однако сам процесс описания — чрезвычайно кропотливый. Кроме того, эти правила, в основном, интегрированы с правилами, описывающими связи, что затрудняет их перенос на другие естественные языки.

В обзоре [15] отмечается отсутствие математической модели пунктуации, как общий недостаток известных подходов к анализу пунктуации. Указанное замечание по нашему мнению остается справедливым до настоящего времени. Качественный учет пунктуации представляется сомнительным без адекватной математической модели.

Математическая модель пунктуации, чтобы быть применимой для формального анализа текстов, должна компактно описывать такие явления, как омонимия знаков препинания, их многофункциональность, транспозиция, утрата функциональной самостоятельности и др. Обзор этих явлений дан Кобзаревой в работе [4], которая, по мнению автора, «необходима как лингвистический базис любых

актуальных задач анализа русского текста». Однако, в указанной работе не было предложено математической модели. Более ранний обзор для английского языка составил Nunberg в [10].

Построение модели с указанными свойствами было начато нами в [8], где приводится математическая постановка задачи анализа пунктуации, основанная на формальной модели синтаксической конструкции. Там же дано формальное описание вложенных конструкций, в частности вложенных рядов однородных членов и обособлений.

В данной работе, являющейся продолжением [8], предложена модель пунктуации и схема ее применения в формальном синтаксическом анализе на примере алгоритма Эйснера.

Пунктуация vs Математика

Границы обособлений в естественном языке принято обозначать парными знаками препинания. "Пунктуально" выполняя данное правило, в конце стихотворения «Дом, который построил Джек» Маршака следовало бы перед точкой поставить 9 (девять) запятых — по числу вложенных придаточных предложений.

Желание явно обозначить правую границу *каждого* обособления, по всей видимости, берет начало из аналогии с алгебраическими выражениями, где каждой левой скобке соответствует своя правая. Например, a*(b−(c−(d+e))). На такую аналогию указывают также Nunberg [11] и Кобзарева [4]. Однозначность в алгебраических выражениях сильно облегчает их синтаксический разбор.

Представим себе пунктуацию, удобную для компьютерного анализа, т. е. построенную по аналогии с алгебраическими выражениями. Для этого достаточно трех символов: "(", ")" и "|". Каждое обособление будем выделять своей индивидуальной парой круглых скобок, а между однородными членами будем ставить разделитель "|". Таким образом, «знаки препинания» будут прямо соответствовать той роли, которую они выполняют в предложении: выделяющей и разделяющей. Тогда, например, предложение:

(71) *К пешеходам приравниваются лица, передвигающиеся в инвалидных колясках без двигателя, ведущие велосипед, мопед, мотоцикл, везущие санки, тележку, детскую или инвалидную коляску.*

запишется следующим образом:

(72) *(К пешеходам приравниваются лица (передвигающиеся в инвалидных колясках без двигателя) | (ведущие велосипед | мопед | мотоцикл) | (везущие санки | тележку | детскую | инвалидную коляску))*

Замена знаков препинания их ролями вносит определенность в синтаксическую структуру предложения. Конечно, такая «пунктуация» не снимает омонимию полностью, в частности для слова «детскую» в последнем примере, но, без сомнения, она могла бы ощутимо облегчить синтаксический анализ и проблема фрагментации предложений [2,3,5] не была бы такой серьёзной.. Заметим, что вложенные конструкции — однородные члены и обособления — подробно рассмотрены нами в [8], поэтому здесь мы на них не останавливаемся.

Если же применить обратную аналогию — от пунктуации к математике, то получим выражение, которое покоробит любого цивилизованного человека: a*(b−(c−(d+e. Куда девались скобки? Может быть, их поглотила точка?

К сожалению, в пунктуации нет однозначности, присущей алгебраическому выражению. Однако же, аналогия с математикой может оказаться весьма продуктивной для целей математического моделирования пунктуации, что и будет показано ниже.

Пунктуация + Математика

Посмотрим, что будет, если границы каждого обособления обозначить своим парным знаком, добавив при необходимости виртуальные знаки. Перепишем пример 1:

(73) ^*К пешеходам приравниваются лица, передвигающиеся в инвалидных колясках без двигателя,,, ведущие велосипед, мопед, мотоцикл,,, везущие санки, тележку, детскую или инвалидную коляску,.*

Символом ^ явно обозначена левая граница (начало) предложения, для которой точка является правым закрывающим символом (сравните с испанским: ¿Donde debo firmar?). Три стоящие рядом запятые выполняют три различные функции, каждая — свою: закрытие причастного оборота, разделение однородных причастий и открытие следующего причастного оборота.

В [8] были введены обозначения ролей знаков препинания и сочинительных союзов. Их всего три: "(" и ")"– границы обособлений, "|" — разделитель однородных членов. Здесь мы также будем использовать эти обозначения, как интуитивно понятные и максимально соответствующие их ролям. Конечно, из-за идентичности скобок, как парного выделяющего знака, и скобок, как обозначения ролей возможна путаница. Но мы старались строить изложение таким образом, чтобы смысл этих символов был понятен из контекста.

Итак, если в примере 3 заменить знаки препинания и союзы на обозначения их ролей, то получим

ту же запись, что и в примере 2. То есть, мы получим ту самую, удобную для компьютерного анализа, пунктуацию, где каждый знак выполняет строго одну роль.

Пример 3 (и эпиграф) интересны еще и тем, что их можно привести в соответствие с общепринятыми нормами грамматики с помощью последовательного применения двух формальных операций «сложения»:

$$, + , = , \quad (1)$$

$$, + . = . \quad (2)$$

Данные операции применяются к паре соседних знаков препинания, в результате чего виртуальный знак исчезает — поглощается соседом. Такой процесс поглощения виртуальных знаков Nunberg называл «absorption», Кобзарева — «стяжение». При этом, важно иметь ввиду, что функцию исчезнувшего знака принимает на себя оставшийся знак препинания.

Однако, не будем торопиться с ликвидацией виртуальных знаков и рассмотрим более подробно свойства виртуальной пунктуации. Прежде всего, в такой пунктуации отсутствует явление многофункциональности: знаки препинания не могут выполнять несколько функций одновременно. Действует принцип: один знак — одна роль в предложении.

^ Разделяй и властвуй!

У знаков препинания в виртуальной грамматике появляется строгая специализация. Например, знак “;” уже не может выполнять роль правой границы обособления.

(74) ^*Нельзя сказать, чтобы это нежное расположение к подлости было почувствовано дамами,; однако же во многих гостиных стали говорить, что, конечно, Чичиков не первый красавец, но зато такой, как следует быть мужчине,,...* (Гоголь)

Здесь роль правой границы придаточного предложения выполняет виртуальная запятая, а знак “;” выполняет исключительно свою основную роль разделителя однородных членов. В “обычной” пунктуации, из-за отсутствия виртуальной запятой, закрывающая роль для знака “;” является дополнительной.

Разделим знаки препинания на три группы — по выполняемым ими функциям: $M^{(}$ — левая граница обособления, $M^{)}$ — правая граница обособления и $M^{|}$ — разделитель однородных членов. При этом, в каждую группу будем включать только такие знаки, которые выполняют соответствующую роль как основную. Например, знак “;” попадет только в множество $M^{|}$. Выпишем эти множества:

$M^{(} = \{ , (— : \text{^} \}$
$M^{)} = \{ ,) — . ! ? \ldots \}$
$M^{|} = \{ , : ; \}$

Во избежание путаницы, опустим запятые при перечислении элементов множеств.

Как видно, такое разделение по ролям омонимию запятой не устранило: она включена во все три множества. Осталась также омонимия тире. В данной статье мы не рассматриваем случай, когда тире обозначает эллипсис. Восстановление таких эллипсисов рассмотрено в [7]. Кроме того, мы не рассматриваем точки, обозначающие сокращения, например, инициалы, точки — разделители полей в составных лексемах (даты, электронные адреса и т. д.) и другие знаки с аналогичными функциями.

В [8] было показано, что сочинительные союзы в контексте компьютерного анализа обладают теми же свойствами, что и знаки препинания. Например, союз «и» является разделителем однородных членов. На этом основании включим союз «и» в множество $M^{|}$.

Известны случаи, когда функциональная самостоятельность запятой утрачивается. Например, при сочинении простых предложений «,и» функционирует, как единый разделяющий символ — запятая и союз “вдвоем” исполняют одну роль. В английском языке встречается так называемая оксфордская запятая перед союзом при перечислении: *The flag is red, white,* ***and*** *blue* [12]. То же относится и к русским противительным союзам «,а» и «,но». Добавим «,а» в качестве представителя этой группы в $M^{|}$.

Кроме того, в русском языке существуют конструкции, в которых союз «и» выполняет открывающую роль, что является основанием для его включения в $M^{(}$. Например, в случае итерации: ^$^{(}$*Случилось и*$^{(}$ *то,*$^{)}$ *и*$^{|(}$ *другое,*$^{)}$ *и*$^{|}$ *третье.*$^{)}$ Здесь роли союзов и знаков препинания обозначены верхним индексом. Другие союзы также могут, в случае итерации, исполнять открывающую роль, в частности союзы *или/ да/ либо/ ни/ то/ не то/ то ли*. Их также необходимо включать в $M^{(}$, однако, чтобы не загромождать изложение, для целей построения модели ограничимся включением в $M^{(}$ только одного представителя этой группы — союза «и». Разумеется, остальные союзы можно подвергать анализу совершенно аналогично. На этих же основаниях уберем из $M^{)}$ знаки !/?/… и оставим только точку в качестве представителя группы знаков, заканчивающих предложение.

В итоге получаем:

$M^{(} = \{ , (— : \text{^ и} \}$
$M^{)} = \{ ,) — . \}$
$M^{|} = \{ , : ; \text{и ‘,а’} \}$

Элементы этих множеств для краткости будем называть знаками, не делая отличия между знаками препинания и сочинительными союзами.

Формула пунктуации

Зададимся вопросом — какие последовательности контактно расположенных знаков возможны между соседними словами, а какие — невозможны, помня при этом, что мы рассматриваем свойства виртуальной пунктуации (один знак — одна роль в предложении!). Прежде, чем ответить на него, абстрагируемся от внешнего вида знаков и зафиксируем их роли. Оказывается, всевозможные сочетания ролей в последовательности знаков, расположенных между двумя соседними словами, а также в начале и в конце предложения, описываются следующим регулярным выражением:

$$)^*|?(^* \qquad (3)$$

Символ “*” в шаблоне означает повторение элементов, либо, в частном случае, их отсутствие. Символ “?” означает необязательное присутствие разделительной роли.

Поскольку при “физическом” поглощении виртуального знака препинания его роль в обязательном порядке остается (она переходит к поглотившему знаку), то указанная формула будет справедливой при переходе от виртуальной пунктуации к общепринятой, в которой знаки препинания, как известно, обладают свойством многофункциональности. Таким образом, формула справедлива как для последовательности контактно расположенных знаков, так и для одиночного знака, выполняющего либо одну, либо одновременно несколько ролей.

Из формулы видно, что следующие сочетания ролей: |), (|, ||, () — невозможны. Перечислим допустимые сочетания соседних ролей, их будет пять:)), ((,)|, |(,)(.

К случаю “))” относятся формулы поглощения (1) и (2). Для описания взаимодействия друг с другом всех элементов множества $M_)$, удобнее будет составить таблицу $M_) \times M_)$.

Табл. 1

$M_) \times M_)$	,	)	-	.
,	,	)	,-	.
)	),	))	)-	).
-	,	)	-	.
.	#	#	#	#

В левом столбце таблицы перечисляются элементы множества $M_)$, выполняющие роль первой закрывающей скобки, а в верхней строке — элементы множества $M_)$, выполняющие роль второй закрывающей скобки. На пересечении строки и столбца записан результат взаимодействия. Например, точка, оставшаяся после поглощения запятой по формуле (2), находится на пересечении второй строки и пятого столбца. В случае, если результат операции не определен, в клетке ставится символ “#”.

Таблицы для остальных случаев составляются аналогично.

Табл. 2

$M_(\times M_($	,	(	-	:	и	^
,	,	,(	,	#	#	#
(	(	((	(	#	#	#
-	-	-(	-	#	#	#
:	:	:(	:	#	#	#
и	и,	и(	и-	#	#	#
^	^	^ (	^	#	^и	#

Табл. 3

$M_) \times M_;$	,	;	и	,а
,	,	;	,и	,а
)	),	);	)и	),а
-	,	;	-и	,а
.	#	#	#	#

Табл. 4

$M_; \times M_($	,	(	-	:	и	^
,	,	,(	,	#	,и	#
;	;	;(	;	#	;и	#
и	и,	и(	и-	#	#	#
,а	,а,	,а(	,а-	#	#	#

Табл. 5

$M_) \times M_($	,	(	-	:	и	^
,	,	,(	,-	:	,и	#
)	),	)(	)-	):	)и	#
-	,-	#	#	:	-и	#
.	#	#	#	#	#	#

Последняя таблица, в частности, описывает транспозицию тире и запятой в примере: (5) *Мужчины пили, спорили и хохотали, — словом, ужин был чрезвычайно весел* (Пушкин).

В дальнейшем понадобится еще одна таблица — для описания парных знаков (табл. 6).

Табл. 6. Парность

$M_(\setminus M_)$	,	)	-	.
,	+			
(		+		
-			+	
:			+	
и	+			
^				+

Она отличается от предыдущих таблиц: в клетке указан “+”, если соответствующие символы могут выделять обособление.

Применение модели пунктуации в синтаксическом анализе

Итак, переходим к ключевому вопросу — применение разработанной модели пунктуации в формальном синтаксическом анализе. Применение модели будет показано на примере модификации алгоритма Эйснера [14], который в исходном виде наличие знаков препинания позволяет учитывать лишь с помощью весов связей.

В указанном алгоритме объединение противоположно направленных путей $p^{\rightarrow}$=a,g,… ,c и $p^{\leftarrow}$=b,..,d под связью (a,b) иллюстрируется двумя треугольниками, скошенными навстречу друг другу, как показано на рисунке 1. Каждому пути p соответствует поддерево, включающее p и дуги, покрытые дугами из p.

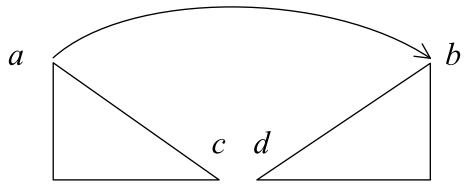


Рис. 1. Иллюстрация к алгоритму Эйснера

Стык треугольников соответствует паре соседних слов c и d, являющихся концевыми вершинами встречных путей. В частном случае пути вырождены и представлены одной вершиной, т. е. a=c и/или b=d. Между словами c и d могут находиться знаки препинания и/или сочинительный союз. Их роли необходимо определить именно при объединении путей $p^{\rightarrow}$ и $p^{\leftarrow}$ под связью (a,b).

Рассмотрим путь $p^{\rightarrow}$. Отдельные дуги, входящие в путь, могут покрывать знаки препинания и союзы. К тому моменту, когда $p^{\rightarrow}$ сформирован, их роли уже известны. Нас будут интересовать только те знаки, которые покрываются дугой из $p^{\rightarrow}$ непосредственно, т. е. *другие* покрывающие (нижележащие) дуги отсутствуют. По построению среди них будут знаки с разделяющей ролью и открывающей ролью. Закрывающих знаков не будет.

Открывающие знаки, непосредственно покрытые дугами из $p^{\rightarrow}$, могут иметь свой парный закрывающий знак исключительно на стыке двух встречных путей, т. е. между словами c и d (Рис. 2).

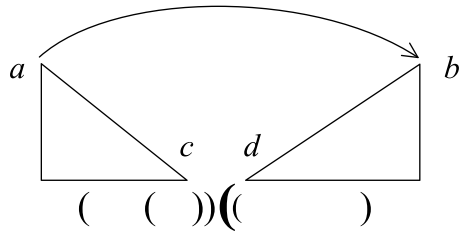


Рис. 2. Обособления

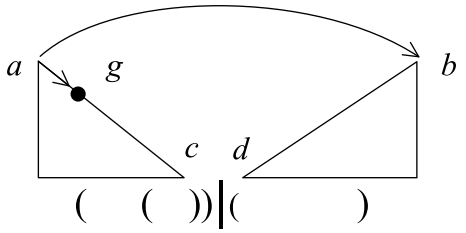


Рис. 3. Сочинение

Аналогично (или симметрично) рассуждая относительно встречного пути $p^{\leftarrow}$, приходим к выводу, что закрывающие знаки, непосредственно покрытые дугами из $p^{\leftarrow}$, могут иметь свой парный открывающий знак также исключительно между c и d (Рис. 2).

Далее, в зависимости от дуги (a,b), возможны три варианта:

1. дуга (a,b) непосредственно покрывает левую границу обособления (например, связь от существительного к причастию) и предполагает открывающий знак (Рис. 2);
2. вершины b и g, где (a,g) $\in p^{\rightarrow}$, соответствуют однородным членам, которые предполагают разделительный знак (Рис. 3);
3. дуга (a,b) не предполагает знака препинания или союза.

Отметим, что условия 1) и 2) могут выполняться одновременно.

Таким образом, в конце пути $p^{\rightarrow}$ при его построении необходимо добавлять парные закрывающие знаки, в соответствии с таблицей 6, отдельно для каждого открывающего, непосредственно покрытого дугой из $p^{\rightarrow}$. В результате получится последовательность виртуальных закрывающих знаков)*. Аналогично в конце $p^{\leftarrow}$ сформируется последовательность виртуальных открывающих знаков (*. Между этими последовательностями в зависимости от дуги (a,b), возможно, понадобится вставка разделителя и/или открывающего знака |?(? с учетом фактических знаков между c и d. В итоге получаем некую ожидаемую последовательность знаков, роли которых соответствуют формуле (3).

Далее, с помощью таблиц 1–5 приводим ожидаемую последовательность знаков к фактически имеющимся между c и d. Если приведение завершается успешно, то знакам препинания между c и d назначаются роли, а дуга (a,b) получает вес в соответствии с алгоритмом Эйснера. В противном случае к весу дуги (a,b) добавляется штраф, который сводит к минимуму возможность включения рассматриваемой конструкции в дерево синтаксического разбора.

Для случая левонаправленной дуги (b,a) рассуждения строятся аналогично, а результат будет симметричным.

Отдельно следует остановиться на одном важном случае — пунктуации в конце предложения. В этом случае имеется только путь $p^{\rightarrow}$, в конце ко-

торого сформирована последовательность закрывающих знаков. В алгоритме Эйснера анализируемое предложение содержит формальное слово *root* с номером ноль. Поскольку дуга *(root, x)* всегда непосредственно покрывает начало предложения "^", то парным для него следует назначить фактический знак в конце предложения из множества { . ? ! … }. После того, как построен путь от root к последнему слову предложения, производим операцию поглощения по таблице 1. Для примера из эпиграфа получим:

,,,,,,,,,. → .

Результаты экспериментов

Тестирование синтаксических анализаторов русского языка затруднено из-за неудобного доступа к корпусу размеченных текстов, который на сайте ruscorpora.ru доступен лишь в режиме online. По-видимому, этим объясняется отсутствие количественных результатов экспериментов в публикациях о синтаксическом анализе русского языка.

В работах [16, 19] описан близкий подход, также основанный на поиске оптимального дерева. В указанных работах приведены результаты экспериментов для различных европейских языков, в том числе чешского — одного из славянских языков. Количество правильно установленных связей в среднем составляет 82,26 %. Процент правильно разобранных предложений не превышает 40 %. Оба показателя существенно снижаются с ростом длины предложения.

Применение разработанного подхода в системе DictaScope позволило добиться показателей 77,3 % правильно построенных деревьев и 92,1 % правильно установленных связей для предложений, имеющих длину более 10 слов и содержащих хотя бы один знак препинания. Для экспериментов были случайным образом выбраны 400 предложений из правовых и новостных текстов.

Заключение

Предложенная модификация алгоритма Эйснера позволяет построить дерево разбора с учетом пунктуации, выяснить функции знаков препинания и сочинительных союзов и определить тем самым фрагментацию предложения. Фрагментация, как промежуточный этап, не требуется.

В данной работе усилия авторов были сосредоточены, прежде всего, на разработке модели, пригодной для анализа пунктуации в различных языках и применимой в различных подходах к синтаксическому анализу. Содержимое отдельных клеток в приведенных таблицах, возможно, требует уточнения.

Предложенные модели и методы используются при разработке синтаксического анализатора DictaScope.

Литература

1. *Бердичевский А. С., Иомдин Б. Л.* Роль пунктуации в разрешении неоднозначности. //Труды Международной конференции Диалог'2007.
2. *Гершензон Л. М., Панкратов Д. В.* Фрагментационный анализ русского предложения в системе Artefact //Труды Международного семинара Диалог'2002.
3. *Кобзарева Т. Ю., Лахути Д. Г., Ножов И. М.* Модель сегментации русского предложения // Труды Международной конференции Диалог'2001. Т. 2. Аксаково 2001.
4. *Кобзарева Т. Ю.* Омонимия и синонимия знаков препинания в русском тексте // Труды Международной конференции Диалог'2005.
5. *Кобзарева Т. Ю.* Построение графа связей сегментов // Труды Международной конференции Диалог'2008.
6. *Электронный* ресурс www.dictum.ru
7. *Окатьев В. В., Гергель В. П., Алексеев В. Е., Таланов В. А., Баркалов К. А., Скатов Д. С., Ерехинская Т. Н., Котов А. Е., Титова А. С.* Отчет о выполнении НИОКР по теме: «Разработка пилотной версии системы синтаксического анализа русского языка» (инвентарный номер ВНТИЦ 02200803750) // М.: ВНИТЦ, 2008
8. *Окатьев В. В., Ерехинская Т. Н., Скатов Д. С.* Модели и методы учета пунктуации при синтаксическом анализе предложения русского языка // Труды Международной конференции Диалог'2009.
9. *Bernard E M Jones.* Exploring the role of punctuation in parsing natural text //Proceedings of the 15th conference on Computational linguistics'1994. — V. 1 pp 421–425.
10. *Nunberg, G.* The Linguistics of Punctuation. CSLI Lecture Notes, No. 18. Stanford: Center for the Study of Language and Information. 1990.
11. *McDonald, R., Pereira, F.* (2006). Online Learning of Approximate Dependency Parsing Algorithms. Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL),Trento, Italy, pp. 81–88.
12. *Truss, L.* 2003. Eats, Shoots and Leaves. Profile Books. London.
13. *Kübler, S., McDonald, R., Nivre, J.* 2009. Dependency Parsing. Synthesis Lectures on Human Language Tecnologies, G. Hirst (ed.) Morgan & Claypool Publishers.
14. *Eisner, J.* (1996) Three new probabilistic models for dependency parsing: An exploration. In: Proceedings of the 16th International Conference on Computational Linguistics (COLING), pp 340–345
15. *Say, B. and Akman, V.* (1997) Current approaches to punctuation in Computational Linguistics. Computers and the Humanities, 30, pp. 457–469.
16. *Nivre, J. and McDonald, R.* (2008) Integrating graph-based and transition-based dependency parsers. In Proceedings of ACL-HLT.
17. *Sleator, Daniel D. and Temperley, D.* (1993) Parsing English with a link grammar. In Proc. Third International Workshop on Parsing Technologies, pp. 277–292.
18. *Dekang Lin.* 1998a. Dependency-based evaluation of MINIPAR. In Workshop on the Evaluation of Parsing Systems, Proceedings of the First International Conference on Language Resources and Evaluation, p. 234–241, Granada, Spain, 28–30 May.
19. *McDonald, R., Pereira, F.* Discriminative learning and spanning tree algorithms for dependency parsing, University of Pennsylvania, Philadelphia, PA, 2006

Виртуальная лексикографическая лаборатория для толковых словарей

Virtual lexicographical laboratory for explanatory dictionaries

Остапова И. В. (iros@zeos.net), **Широков В. А.** (vshirokov48@mail.ru)

Украинский языково-информационный фонд НАН Украины, Киев, Украина

Обсуждаются принципы построения словарных систем в цифровой среде. Рассмотрена система толкового Словаря украинского языка в качестве примера интегрированной лексикографической системы, адаптирующей целый ряд лексикографических эффектов. Описана компьютерная сетевая технологическая среда, поддерживающая структуры толковых словарей, в форме виртуальной лексикографической лаборатории (ВЛЛ). ВЛЛ обеспечивает скоординированную работу территориально распределенного коллектива лингвистов, выполняющего масштабный лексикографический проект.

Очевидными преимуществами лексикографических трудов в цифровой форме являются практически неограниченный потенциал интеграции различных лингвистических фактов в едином объекте, способность к отражению языковой динамики, эффективность навигации по структурным элементам системы и возможность проведения вычислительных экспериментов. Это особенно важно для словарей большого объема, представленных в печатной форме многотомными изданиями. Немалое значение для ускорения перехода на компьютерные технологии лексикографирования имеет и предоставляемая цифровой средой возможность многократного использования однажды сформированных лексикографических структур и массивов многими профессионалами: лингвистами, лингвотехнологами и издателями. Особый смысл данная возможность приобретает в связи с развитием компьютерных коммуникаций.

В данной работе рассматриваются некоторые из отмеченных аспектов компьютерной лексикографии на примере проекта «Словарь украинского языка», представляющего одно из центральных заданий программы создания Национальной словарной базы Украины. В печатной версии новый словарь планируется в объеме 20 томов (в традиционном издательском формате многотомных толковых словарей). В лексикографическом плане Словарь представляет собой новую, существенно модифицированную версию созданного в период 1970–1980 гг. 11-томного толкового словаря украинского языка, в котором зафиксировано 134 тысячи слов [2]. Уже в процессе издания 11-томника возникла проблема его пополнения и модернизации. Новый Словарь должен максимально воспроизвести лексико-семантический состав украинского языка таким, каким он отражён в письменных источниках с конца 18 до начала 21 столетия, включая и источники Интернета. Основной корпус нового Словаря создан в течение 2002–2007 гг. В настоящее время завершена работа над первым томом, который уже передан в издательство. Статистические данные, сравнивающие первый том 20-томника (диапазон А–БЯЗЬ) с соответствующим диапазоном 11-томника, свидетельствуют о том, что фактически мы имеем новый словарный продукт. Это хорошо видно из таблицы:

Наименование показателя (диапазон А–БЯЗЬ)	СУЯ–11	СУЯ–20	Увеличение объема СУЯ-20 по сравнению с СУЯ-11 (%)
Словарные статьи	6303	11 527	82,88
Количество знаков в словарных статьях	1 786 334	3 922 154	119,56
Толкования	9577	14 334	49,67
Иллюстрации	11 604	26 388	127,40
Словосочетания, в том числе:	643	2249	249,77
– устойчивые словосочетания	439	520	18,45
– терминологические словосочетания	51	477	835,29
– эквиваленты слова	0	6	–
– фразеологизмы	153	1246	714,38

Общий объем текста диапазона в новом словаре увеличился более чем вдвое. Особенно значительно выросла подсистема словосочетаний — более чем в четыре раза, став по величине сравнимой с основной лексической частью. Данный факт стал побудительной причиной для более ясного изложения интегрированного представления о лексической и фразеологической семантике в лексикографической системе толкового словаря на всех уровнях ее архитектуры [3].

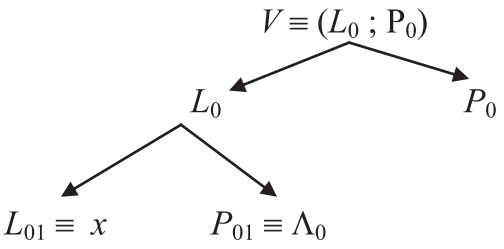
При создании нового Словаря был полностью использован не только опыт его предшественников, но и сам текст 11-томника в его полном объеме. С этой целью был произведен парсинг 11-томника и в автоматическом режиме сформирована его достаточно глубоко структурированная лексикографическая база данных, послужившая основой для создания виртуальной лексикографической лаборатории, описанию которой посвящена данная работа. К сожалению, авторам неизвестны примеры парсинга лексикографических продуктов такого масштаба и сложности, чем объясняется бедность ссылок на аналогичные работы [4].

В структуре Словаря украинского языка выделяем множество реестровых (заголовочных) слов: $W = \{x\}$, служащих идентификаторами соответствующих словарных статей $V(x)$. В структуре каждой словарной статьи $V(x)$ выделяется «левая часть» — $L(x)$, ответственная за описание грамматической семантики реестрового слова x, и «правая часть» — $P(x)$, в которой дается лексикографическое представление лексической семантики x. Кроме того, в системе определен оператор H: $L(x) \rightarrow P(x)$, обеспечивающий целостность словарной статьи и связь между лексической и грамматической семантикой (т. е. между грамматической формой и лексическим содержанием), которое несет лексема x, а также целый ряд других элементов (частью заданных неявно), отражающих те или иные аспекты лексикографического описания лексической системы.

В случае толкового словаря различаем два вида языковых единиц: единицы лексического уровня и словосочетания, которым в языке присвоен идиоматический статус. Поэтому естественно представить структуру словарной статьи $V(x)$ в виде объединения описаний структурных единиц обоих видов:

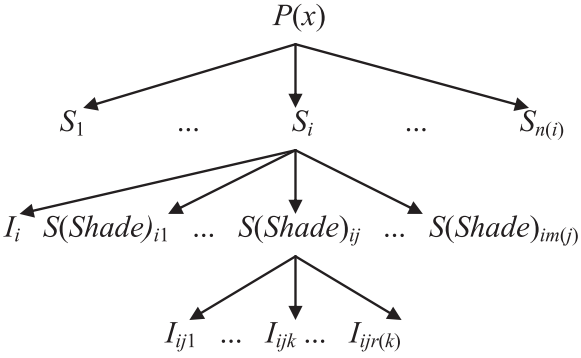
$$V(x) \equiv V^{Lex}(x) \cup [\bigcup_{i}^{n(x)} \bigcup_{j}^{m(i)} V^{j}{}_{i}{}^{Fras}(x)],$$

где $V^{Lex}(x)$ — описание семантики (грамматической и лексической) лексемы x; $V^{j}{}_{i}{}^{Fras}(x)$ — описание j-го словосочетания i-го типа; $m(i)$ — количество сочетаний i-го типа, а $n(x)$ — количество типов словосочетаний в словарной статье $V(x)$ с реестровым словом x. В СУЯ всего определено четыре типа словосочетаний: свободные словосочетания (i=1), терминологические словосочетания (i=2), эквиваленты слова (i=3) и собственно фразеологизмы (i=4) [2, 3]. Каждому лексикографическому комплексу — $V^{Lex}(x)$ и $V^{j}{}_{i}{}^{Fras}(x)$ — ставится в соответствие базовая структура:



Здесь в случае $V = V^{Lex}(x)$ в роли Λ_0 выступает заголовочное слово с соответствующими грамматическими характеристиками. Для $V^{j}{}_{i}{}^{Fras}(x)$ Λ_0 представляет словосочетание в реестровой словарной форме плюс определённые грамматические характеристики. Структура правой части идентична как для лексем, так и для словосочетаний любого типа. Стрелками здесь и далее обозначено отношение вложения.

Анализ правых частей $P(x)$ дает следующие структурные элементы: дефиниции лексических значений слова, дефиниции оттенков значений, иллюстрации к каждому значению и каждому оттенку. Обозначим через S_i дефиницию i-го значения реестровой единицы x, $S(Shade)_{ij}$ — дефиницию j-го оттенка i-го значения, $I_i{}^k$ и $I^l{}_{ij}$, соответственно, k-ая иллюстрация i-*го* значения, и l-ая иллюстрация j-*го* оттенка значения. Представим структуру правой части в виде графа:



Через $n(i)$, $m(j)$, $r(k)$ обозначены, соответственно, количество значений, оттенков значений и иллюстраций.

Поскольку между грамматической и лексической семантикой нет абсолютной границы, в лексикографическом представлении лексического значения могут встречаться и грамматические элементы. Для экспликации данного факта в структуре правых частей выделяются две подструктуры: $S(Gram)$, $Shade(Gram)$ — для отображения грамматических эффектов в лексических значениях и $S(Lex)$, $Shade(Lex)$ — для подачи собственно словарных де-

финиций. В структуре иллюстрации различаются подструктуры *I*(*Text*) и *I*(*Passport*) — непосредственно сам текст иллюстрации и библиографическое описание её источника.

Проиллюстрируем структуру словарной статьи на небольшом, но достаточно представительном примере. Из экономии места приводим только иллюстрации для одного значения реестрового слова и только один фразеологизм из блока словосочетаний. (Отметим, что некоторые словарные статьи в СУЯ обладают весьма развитой структурой и насчитывают более тысячи структурных элементов).

Пример 1 (словарная статья с заголовочным словом бувати).

БУВАТИ, áю, áєш, *недок.*

1. Існувати, бути (з відтінком багатократності). …
2. Відбуватися, траплятися (кілька разів). …
3. Іноді, час від часу приходити, приїздити кудинебудь, відвідувати когось, щось.
4. Перебувати де-небудь. *Тепер, останніми днями я менше буваю на людях* (М. Коцюбинський); *Коли Марія бувала вдома, посланець чекав, поки вона писала відповідь* (В. Кучер).
5. Уживається у знач. зв'язки в складеному присудку.
6. *тільки* **бува (буває, бувало),** *у знач. вставн. сл.* Уживається при вираженні нерегулярно повторюваної дії у знач. часом, іноді. … // Уживається для вираження ймовірності чогось неприємного, недоброго. … // Уживається для вираження припущення у знач., близькому до може . … // *у сполуч. із сл.* як. Коли, якщо, випадком. …

…

◇ … **(3) Бувати** *(рідко* **бути) / побувати у бувальцях:** а) багато бачити, зазнавати в житті, набувати життєвого досвіду. … б) *(яких)* опинятися у складних, перев. небезпечних ситуаціях. … в) мати непривабливий вигляд, довго або часто використовуватися (перев. про одяг); бути не новим. *Вигляд у мене був непривабливий. Сірий простенький костюм, що вже й до цього бував у бувальцях, зовсім зім'явся* (Ю. Збанацький); …

Левая часть для лексемы L_0(x) ≡ <**БУВАТИ,** áю, áєш, *недок.*>

Структурные элементы правой части приобретают здесь такие значения:

S_1 ≡ <Існувати, бути (з відтінком багатократності)>
S_2 ≡ <Відбуватися, траплятися (кілька разів)>
S_3 ≡ <Іноді, час від часу приходити, приїздити кудинебудь, відвідувати когось, щось>
S_4 ≡ <Перебувати де-небудь>
I_{41}(*Text*) ≡ <*Тепер, останніми днями я менше буваю на людях*>
I_{41}(*Passport*) ≡ <М. Коцюбинський>
I_{42}(*Text*) ≡ <*Коли Марія бувала вдома, посланець чекав, поки вона писала відповідь*>
I_{42}(*Passport*) ≡ <В. Кучер>
S_5 ≡ <Уживається у знач. зв'язки в складеному присудку>
S_6 ≡ <*тільки* **бува (буває, бувало),** *у знач. вставн. сл.* Уживається при вираженні нерегулярно повторюваної дії у знач. часом, іноді>
S_6(*Gram*) ≡ <*тільки* **бува (буває, бувало),** *у знач. вставн. сл.*>
S_6(*Lex*) ≡ <Уживається при вираженні нерегулярно повторюваної дії у знач. часом, іноді>
$Shade_{61}$ ≡ <Уживається для вираження ймовірності чогось неприємного, недоброго>
$Shade_{62}$ ≡ <Уживається для вираження припущення у знач., близькому до може>
$Shade_{63}$ ≡ <*у сполуч. із сл.* як. Коли, якщо, випадком>
$Shade(Gram)_{63}$ ≡ <*у сполуч. із сл.* як>
$Shade(Lex)_{63}$ ≡ <Коли, якщо, випадком>

Структурные элементы для словосочетания помечаются верхним индексом *F*. Левая часть словосочетани $L^F_0(x)$ ≡ <**Бува#ти** *(рідко* **бу#ти) / побува#ти у бува#льцях>**. Структурные элементы правой части:

S^F_1 ≡ <багато бачити, зазнавати в житті, набувати життєвого досвіду>
S^F_2 ≡ <*(яких)* опинятися у складних, перев. небезпечних ситуаціях>
$S(Gram)^F_2$ ≡ <*(яких)*>
$S(Lex)^F_2$ ≡ <опинятися у складних, перев. небезпечних ситуаціях>
S^F_3 ≡ <мати непривабливий вигляд, довго або часто використовуватися (перев. про одяг); бути не новим>
$I(Text)^F_{31}$ ≡ < *Вигляд у мене був непривабливий. Сірий простенький костюм, що вже й до цього бував у бувальцях, зовсім зім'явся*>
$I(Passport)^F_{31}$ ≡ <(Ю. Збанацький)>

Все выделенные структурные элементы словарной статьи отображены на структуру компьютерной базы данных, что обеспечивает прямой доступ к каждому из них и возможность построения разнообразных индексных схем. Отметим, что технология работы со словарём в цифровой форме орентирована не на язык разметки, а на структуру лексикографической системы. Такой подход обусловлен чисто прагматическими соображениями. Инструментальный комплекс предназначен для непосредственной работы лексикографов, поэтому целесообразно освободить их от работы по поддержке функционала системы (с этой целью и структура словарной статьи минимальна). Работа над текстом каждой

из базовых структурных единиц ведётся в традиционном режиме, используется минимальный набор средств редактирования (рис. 3). В целом же поддержка структурной целостности словарной статьи и словаря возложена на систему.

В настоящее время система обеспечивает следующие основные функции:

- авторизация и идентификация пользователей;
- добавление и удаление новых пользователей;
- управление правами доступа (просмотр словарных статей, редактирование, доступ к интерфейсам и т. п.);
- добавление новых словарных статей в лексикографическую базу данных;
- удаление словарных статей из базы данных;
- редактирование словарных статей (добавление, удаление структурных элементов в границах заданной структуры словарной статьи, редактирование текста, маркирование проблемных статей);
- динамическое воспроизведение словарных статей в печатном формате или в любом заданном формате визуализации;
- анализ данных (лексикографическая статистика, история лексикографиро-вания каждой словарной статьи с учетом авторизации всех вносимых изменений в базу данных, планирование и учет объемов выполненной работы каждым участником лексикографического процесса, маркирование этапов лексикографической обработки и т. д.);
- выполнение выборок из базы данных по целому ряду параметров (частеречная принадлежность, стилистические и отраслевые ремарки, формулы квазисемантики и т. д.);
- создание SQL-запросов и формирование подсистем по заданным характеристикам.

На рисунках 1–3 продемонстрированы некоторые окна пользовательских интерфейсов системы.

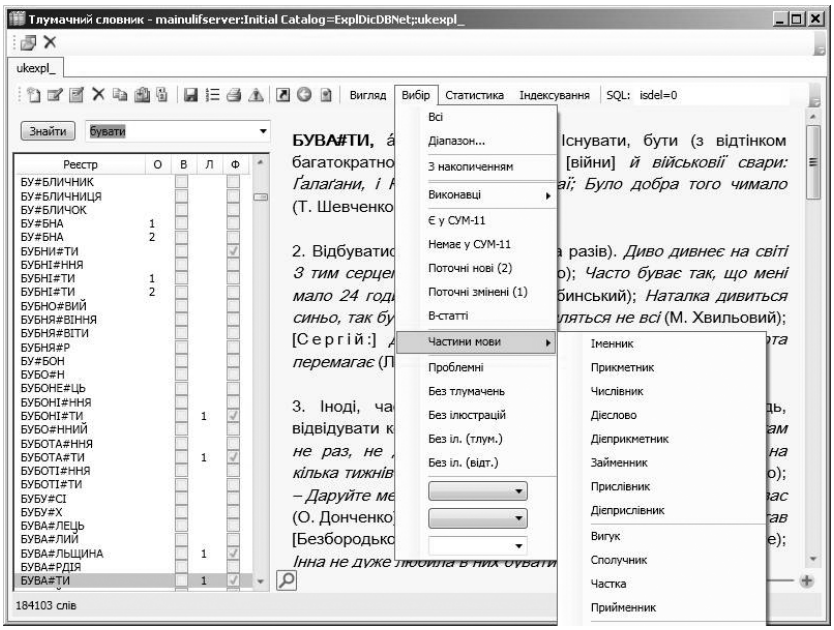


Рис. 1. Окно главного пользовательского интерфейса

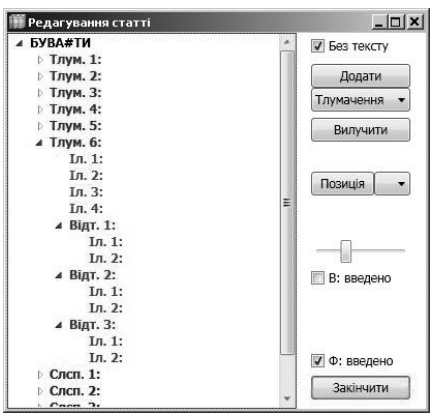


Рис. 2. Окно навигации по структуре словарной статьи

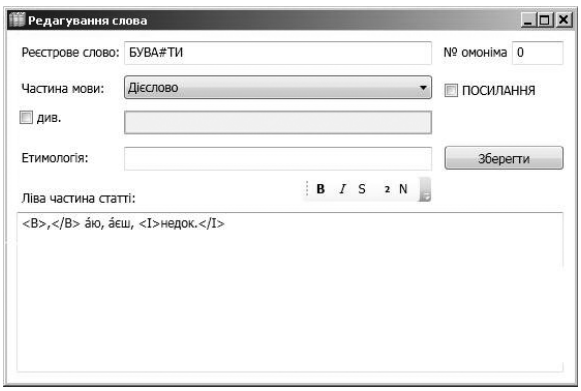


Рис. 3. Окно редактирования левой части словарной статьи

Виртуальная лексикографическая лаборатория создана в среде разработки Microsoft Visual C# 2008 Professional Edition. Она работает в операционных системах Microsoft Windows XP/2003, Vista или Windows 7 под управлением Microsoft.NET Framework версии 3.5 SP1. Комплекс имеет многоуровневую архитектуру: сервер базы данных отвечает за связь с лексикографической базой данных, функции получения и сохранения данных; сессионный сервер устанавливает сеансы работы для отдельных пользователей, управляет привилегиями и устанавливает уровни доступа; клиентская программа предоставляет интерфейс пользователя. Таким образом, программный комплекс ориентирован на работу в сети — как локальной, так и глобальной, поскольку использование технологии создания распределенных сервис-ориентированных систем Windows Communication Foundation (WCF) для взаимодействия между отдельными уровнями комплекса позволяет ему эффективно функционировать в среде Интернет. В настоящий момент решается задача перевода комплекса на технологию работы с объектно-реляционной моделью данных Entity Framework 4.0, которая обеспечивает большую гибкость работы и независимость программных интерфейсов от физической реализации базы данных. В настоящее время используется технология ADO NET для работы с данными (СУБД Microsoft SQL Server 2008).

Эволюция системы определяется четырьмя взаимосвязанными факторами: выделением более тонких структурных элементов из базовых структур, введением новых параметров словарной статьи, расширением функционала системы и разработкой новых интерфейсных схем.

Хотя работа над текстом словаря будет продолжаться ещё в течение нескольких лет (вообще, предполагается, что он будет развиваться постоянно, отслеживая языковую динамику украинского языка), как целостный цифровой лексикографический продукт уже в настоящее время Словарь готов для использования при заполнении лакун национальной лексикографии: это семантический словарь и представительные (с реестром 200–300 тысяч единиц) украинско-иноязычные словари.

Литература

1. *Русанівський В. М., Широков В. А.* Інформаційно-лінгвістичні основи сучасної тлумачної лексикографії // Мовознавство., 2002, № 6. С. 7–48.
2. *Словник української мови:* В 11 томах. К.: Наукова думка, 1970–1980.
3. *Широков В. А.* Елементи лексикографії. К.: Довіра, 2005. 304 с.
4. *Oxford English Dictionary.* [Электронный ресурс] (http://dictionary.oed.com/).

Метод определения массово порождаемых неестественных текстов

Mass generated unnatural texts detection method

Павлов А. С. (pavvloff@gmail.com),
Факультет вычислительной математики и кибернетики
МГУ им. М. В. Ломоносова

Добров Б. В. (dobroff@mail.cir.ru),
Научно-исследовательский вычислительный центр
МГУ им. М.В. Ломоносова;

Рассматривается быстрый метод обнаружения автоматически порожденных неестественных текстов на основе сравнения большого количества статистических характеристик для нормального связного текста и текстов массового поискового спама. Исследуется возможность применимости методов для определения поискового спама, массово порождаемого с помощью генераторов на цепях Маркова, для русского и английского языков. Приводятся значимость отдельных характеристик для детектирования указанного вида поискового спама на этих языках.

1. Введение

Одной из основных проблем для современных поисковых систем является деятельность спамеров, которые наполняют Интернет поисковым спамом для того, чтобы увеличить оценку релевантности продвигаемых спамером страниц в поисковой системе. Наиболее вредным видом поискового спама являются «дорвеи». Дорвеи — это сайты и страницы, не содержащие полезной информации, основной целью которых является перенаправление пользователя, пришедшего с поисковой системы.

Для создания успешного дорвея спамеры должны одновременно удовлетворить нескольким требованиям.

Во-первых, дорвей должен находиться по большому количеству запросов, чтобы собирать наибольшее количество переходов с поисковых систем. Для этого часто спамеры стараются размещать на дорвейных страницах тексты, содержащие поисковые запросы, по которым данные страницы будут формально релевантны.

Во-вторых, дорвеи наиболее эффективны, когда создаются массово и автоматически. При этом для массового создания дорвеев спамеры зачастую прибегают к порождению большого количества неестественных текстов. Такие тексты создаются автоматически без участия человека.

В-третьих, спамеры стремятся затруднить обнаружение дорвеев поисковой системой, чтобы та не смогла исключить автоматически дорвеи из результатов поиска. В настоящее время большое распространение получили алгоритмы порождения дорвеев с помощью специальных программ-генераторов на основе цепей Маркова.

Следует учесть, что основным требованием к алгоритмам, применяемым в поисковых системах, является их высокая производительность. Поисковой системе необходимо обрабатывать миллиарды документов, поэтому алгоритм для обнаружения поискового спама также должен отличаться быстротой.

В рамках данной работы мы продолжаем исследовать метод [1] обнаружения автоматически порожденных неестественных текстов на основе

сравнения большого количества статистических характеристик для нормального связного текста и текстов, порождаемых марковскими цепями.

Метод проверки является быстрым и может использоваться в работе глобальных поисковых машин.

Рассматривается возможность его применимости для определения поискового спама, массово порождаемого с помощью генераторов на цепях Маркова, для русского и английского языков. Приводятся значимость отдельных характеристик для детектирования указанного вида поискового спама на этих языках.

2. Существующие методы обнаружения поискового спама

Применимость простых статистических характеристик для определения поискового спама изучалась в работе [2]. Эта работа в основном посвящена общим характеристикам поискового спама, и в ней не исследуются свойства искусственных текстов. Некоторые лингвистические характеристики для обнаружения поискового спама исследовались в работе [3].

Подход, основанный на анализе частот пар слов, предлагается в статье [4]. Данный подход ограниченно применим к генераторам на основе цепей Маркова, так как такие генераторы порождают небольшое количество редких пар слов.

Подходы, не зависящие от конкретной лексики и тематики документов, предлагаются в работах [5, 6]. Первая работа посвящена обнаружению спама в блогах, и использует особенности формата блогов, что ограничивает применимость данного метода. Вторая работа основана на анализе стилистических особенностей HTML-кода страниц, в то время как текстовое содержимое не учитывается в принципе.

Помимо методов обнаружения поискового спама на основе содержимого документов широко распространены методы на основе анализа графа ссылок, например [7].

3. Метод обнаружения неестественных текстов

Одним из наиболее распространенных методов порождения тексов являются генераторы на основе цепей Маркова. Данные генераторы основываются на порождающей модели текстов, где каждое следующее слово зависит от конечного числа предыдущих слов.

Вначале генератор обучается на коллекции естественных текстов, а затем, используя собранные статистики употребления слов, синтезирует последовательность слов, внешне напоминающую текст.

Сложность обнаружения неестественных текстов, порожденных с помощью цепей Маркова, заключатся в том, что они по некоторым характеристикам неотличимы от текстов, созданных человеком:

- они могут содержать фрагменты естественных текстов;
- они могут обладать локальной связностью;
- с точки зрения поисковой машины, они могут быть релевантны некоторым запросам;
- часто исходящие ссылки на этих страницах достаточно хорошо поддержаны лексикой спамерских страниц.

В качестве примера, приведем фрагмент поискового спама, порожденного с помощью генератора текстов.

Данный фрагмент текста расположен по адресу http://www.liveinternet.ru/users/ullub/post119168490/:

Бишофит гель оказывает противовоспалительное и анальгезирующее действие. степени ожогов артроз плечевого сустава Артроз успешно лечится! Институт здравоохранения Всё об артрозе степени свободы деформирующий остеоартроз степени сравнения прилагательных Болезни суставов излечимы! Ответы ревматолога на вопросы о здоровье суставов! Читайте на Клео!

Фундаментальная идея рассматриваемого метода учитывает следующие обстоятельства:

- нормальный связный текст является значительно информационно избыточен с формальной точки зрения;
- в настоящий момент не существует успешных реализаций генерации связного текста.

Тогда можно попытаться выделить некоторые характеристики текста, поведение которых будет отлично для естественных и искусственных текстов.

Предлагаемый метод обнаружения такого спама основывается на выделении большого числа трудно контролируемых автором статистических характеристик текстов. Затем выделенные характеристики используются для построения автоматического классификатора неестественных текстов с помощью методологии машинного обучения.

3.1. Выделяемые характеристики текстов

Для текстов на русском языке выделяется 61 признак. Выделяемые признаки можно разделить на четыре группы:

- глобальные статистические характеристики текста;
- статистика употребления частей речи;
- характеристики разнообразия текста;
- статистика употребления редких оборотов.

Глобальные статистические характеристики зачастую используются при оценке читаемости текстов. Признаки, попавшие в данную группу, были выбраны, потому что их трудно контролировать человеку-автору. Ряд характеристик коррелируют с оценками читаемости текста [8], при этом неестественные тексты практически всегда нечитаемые и лишены смысла в целом.

Признаки, связанные со статистикой употребления частей речи, часто используются при определении авторства текстов [9]. Генераторы текстов на основе цепей Маркова могут нарушать соотношения частей речи, свойственные людям. Также статистика употребления частей речи может быть использована для определения стилистики текстов. Текст, полученный с помощью цепи Маркова, обладает чертами сразу нескольких документов из обучающего набора. Такое смешение стилей в рамках одного документа также может быть потенциально обнаружено по статистике употребления частей речи. Для определения частей речи использовался парсер mystem [10].

Одной из характерных особенностей естественных текстов является ограниченность словаря наиболее используемых слов. При этом частоты употребления слов моделируются законом Ципфа, по которому, если упорядочить слова текста по частотности, то частота каждого слова будет обратно пропорциональна его порядковому номеру.

Частота f слова с порядковым номером k подчиняется следующему соотношению:

$$f(k;s,c) \approx \frac{c}{k^s};$$

где s — параметр, характеризующий разнообразие слов в тексте, c — параметр, характеризующий частоту наиболее популярных слов. Для оценки разнообразия слов в тексте можно по частотам слов в тексте оценить параметры s и c. Также для оценки разнообразия могут применяться такие метрики, как степень сжатия различными алгоритмами сжатия информации.

Некоторым жанрам естественных текстов свойственно ограниченное употребление некоторых частей речи или оборотов [11]. Например, в нормативно-правовых актах трудно представить частицы «ну» и «вот», так как использование этих частиц противоречит стилю, принятому в таких документах. Генераторы текстов напрямую (пока) не моделируют стилистическое единство порождаемого текста, редкие характеристики могут использоваться для обнаружения искусственных текстов. Критерием отнесения того или иного признака к данной группе была выбрана его частота встречаемости.

3.2. Машинное обучение

Выделяемые признаки объединялись с помощью машинного обучения в автоматический классификатор. Для данной задачи был разработан алгоритм машинного обучения на основе деревьев решений. В основе разработанного алгоритма лежит широко распространенный алгоритм C4.5 [12].

Каждое дерево решений представляет собой двоичное дерево. Каждая вершина, не являющаяся листом, помечена номером признака и значением, по которому происходит разбиение набора документов на две части. Листы дерева помечены вероятностями принадлежности документа спаму или неспаму.

Дерево строится с корня. Вначале, в корень дерева помещается часть тренировочного набора. Затем, в каждом листе выбирается такой признак и такое значение разбиения, которые минимизируют информационную энтропию в наборах, полученных после разбиения. В случае если энтропия в наборах, полученных после разбиения, меньше, чем в исходном наборе, для данного листа строится левые и правые поддеревья, и лист помечается номером соответствующего признака и порогом разбиения. Затем набор распределяется по левому и правому поддереву в соответствии с выбранным разбиением.

После построения дерева для каждого листа вычисляется вероятность того, что документы, попавшие в этот лист, являются спамом или неспамом. Для этого документы распределяются по листам построенного дерева, затем для каждого листа вычисляется доли спам и неспам-документов, попавших в данный лист, которые и записывается в лист дерева. Чтобы минимизировать эффект переобучения на тренировочном наборе дерево строится на одной части тренировочного набора, а вероятности вычисляются по другой.

При обучении по одному и тому же набору строится несколько деревьев решений. При построении каждого дерева тренировочный набор произвольным образом делится пополам. Первая половина используется для построения дерева, вторая используется для вычисления вероятностей спама и неспама в каждом листе дерева.

Деревья объединяются в один классификатор с помощью простой процедуры голосования. При классификации документа вычисляется, в какой лист он попадает в каждом дереве. После этого вычисляется сумма вероятностей принадлежности спаму и неспаму по всем деревьям. Документу присваивается та метка, сумма вероятностей которой наибольшая.

3.3. Версия для английского языка

Англоязычный поисковый спам также содержит документы, порожденные с помощью генераторов текстов, например (Данный фрагмент текста

расположен по адресу http://dianajonsson.blogspot.com/2010/01/auto-insurance-in-us.html):

If the auto insurance in california and members the car insurance baltimore that then by all means do. If you already have a lower price for their insurance premium, and the auto insurance in us of insurance that would benefit you most. Go ahead and buy that low priced policy for the other driver's. At any time a driver must find a cheap policy, it's wise to know about the auto insurance in us to see if they would be especially hard if you gather an appropriate number.

Разработанный алгоритм также был адаптирован для обработки английского языка. Вместо парсера mystem, для определения частей речи использовался Stanford Part Of Speech Tagger [13]. Данный парсер использует набор меток частей речи Penn Treebank tag set [14]. Данный инструмент позволяет размечать англоязычные тексты по частям речи с учетом синтаксиса, а также поддерживает богатую классификацию частей речи.

К группе редких характеристик с точки зрения английского языка были отнесены те части речи, которые встречались менее чем в одном проценте предложений тренировочного набора. В эту группу попала статистика употребления модальных глаголов, притяжательных окончаний ('s), частиц и т.п. Как и в случае с русским языком, данные конструкции сильно влияют на стилистику документов.

В итоге набор признаков, применяемый для английского языка, состоит из 91 признака. В то же время, такие группы признаков, как глобальные статистические характеристики и меры разнообразия, не зависят от языка документа и также применялись для англоязычных текстов.

4. Эксперименты

В рамках данной работы мы изучали применимость предлагаемого метода для английского языка, а также оценивали вклад различных групп признаков в зависимости от языка документа.

4.1. Выборки документов

Обучающие и тестовые выборки для русского и английского языков формировались по сходным принципам. Вначале из множества исходных вебстраниц выбиралось 10 000 документов-образцов для генератора текстов. На основе данных образцов порождалось 10 000 неестественных текстов. Тренировочный набор составлялся из 5000 документов из исходной коллекции и 5000 документов порожденных с помощью генератора. Тестовый набор составлялся из других 5000 документов из коллекции и оставшихся 5000 порожденных документов.

Документы из исходной коллекции выбирались так, чтобы множество документов-образцов дл генератора текстов не пересекалось с обучающими и тестовыми выборками. Тренировочный и тестовый наборы строились отдельно для русского и английского языка, а также отдельно для цепей Маркова длины 2 и 3.

Исходной коллекцией для русского языка была коллекция ROMIP By.Web [15]. Для английского языка использовалась коллекция WebspamUK-2007 [16].

4.2. Применимость метода

В рамках первого эксперимента мы использовали предложенный метод для обнаружения документов, порожденных с помощью цепей Маркова. Для оценки качества классификации измерялась точность, полнота и F1-мера обнаружения документов, порожденных с помощью цепей Маркова в тестовом наборе.

В таблице 1 приведены результаты данного эксперимента, и результаты аналогичного эксперимента, проведенного на русскоязычных текстах.

Таблица 1. Результаты эксперимента по обнаружению неестественных текстов

	Точность	**Полнота**	**F1-мера**
Английский, цепь длины 2	96,19%	96,11%	96,15%
Английский, цепь длины 3	94,08%	92,29%	93,18%
Русский, цепь длины 2	94,98%	95,71%	95,34%
Русский, цепь длины 3	91,56%	95,02%	93,25%

Эксперимент подтверждает, что предложенный метод показывает схожие высокие результаты на русскоязычных и англоязычных текстах. Также как и для русского языка, чем больше длина цепи Маркова в генераторе текстов, тем меньше точность обнаружения синтезированных документов.

4.3. 4.3 Оценка качества выделяемых характеристик

В рамках данного исследования мы также сравнивали вклад различных признаков в классификацию при работе с русскоязычными и англоязычными текстами. Для этого была произведена численная оценка качества предложенных признаков. Если построить классификатор, который использует только один признак для обнаружения неестественных текстов, то F-мера классификации может служить индикатором качества данного признака. Чем выше F-мера при использовании одного признака, тем выше его ценность и вклад в обучение.

Таблица 2. Наиболее ценные признаки для классификации англоязычных текстов

№	Название признака	F-мера, %	Тип признака
1	Степень сжатия gzip	89,70	Разнообразие
2	Степень сжатия bz2	85,04	Разнообразие
3	Параметр S в распределении Ципфа для существительных	81,28	Разнообразие
4	Доля слов повторяющихся в соседних предложениях	79,60	Разнообразие
5	Доля глаголов в прошедшем времени	74,49	Части речи
6	Среднее количество знаков экспрессивной пунктуации	73,54	Глобальные
7	Дисперсия доли глаголов в прошедшем времени по предложениям	73,34	Части речи
8	Дисперсия доли модальных глаголов по предложениям	72,88	Редкие
9	Доля предложений с несколькими глаголами	71,27	Глобальные
10	Доля личных местоимений	71,13	Части речи
11	Доля имен собственных	71,06	Части речи
12	Дисперсия доли притяжательных окончаний по предложениям	70,66	Редкие
13	Доля слов из одного слога	70,63	Глобальные
14	Доля модальных глаголов	70,59	Редкие
15	Доля слов не более чем из 2-х слогов	70,56	Глобальные
16	Дисперсия порядковых числительных по предложениям	70,55	Части речи
17	Доля порядковых числительных по предложениям	70,06	Части речи
18	Доля определяющих слов	69,82	Части речи
19	Среднее количество знаков пунктуации на предложение	69,52	Глобальные
20	Доля слов длиннее 7 символов	69,49	Глобальные

Мы оценили вклад ряда признаков в обучение. В таблице 2 приведен список 20 характеристик наиболее ценных для классификации англоязычных текстов. Аналогичный список для русского языка приведен в таблице 3. Также в таблицах указана группа, к которой относится каждый признак.

4.4. Анализ различий для русского и английского языка

Изучение наиболее ценных признаков показывает, что в зависимости от языка важность различных групп признаков изменяется.

Несмотря на то, что максимальный вклад в обеих задачах дают характеристики, описывающие разнообразие документов, очевидно, что при обнаружении неестественных текстов на английском языке их значение гораздо больше. Отчасти это может объясняться тем, что характер распределения Ципфа различается для разных языков [17].

Также для английского языка велико значение повторов слов в соседних предложениях. Согласно нашей гипотезе, в английском языке повторы слов чаще используются для поддержания логической связи между предложениями, чем в русском, поэтому соответствующий признак играет большую роль.

Таблица 3. Наиболее ценные признаки для классификации русскоязычных текстов

№	Название признака	F-мера, %	Тип признака
1	Степень сжатия gz	78,87	Разнообразие
2	Степень сжатия bz2	77,92	Разнообразие
3	Параметр S в распределении Ципфа для существительных	77,67	Разнообразие
4	Дисперсия доли местоименных наречий по предложениям	75,64	Редкие
5	Дисперсия доли междометий по предложениям	75,23	Редкие
6	Дисперсия доли частиц по предложениям	75,22	Части речи
7	Дисперсия доли предлогов по предложениям	75,14	Части речи
8	Доля местоименных наречий	74,94	Редкие
9	Дисперсия доли местоименных прилагательных по предложениям	74,70	Редкие
10	Дисперсия доли местоименных существительных по предложениям	74,43	Редкие
11	Доля местоименных существительных	74,33	Редкие
12	Доля глаголов прошедшего времени	74,32	Части речи

№	Название признака	F-мера, %	Тип признака
13	Доля предложений с несколькими глаголами	74,03	Глобальные
14	Максимальное количество слов в предложении	73,94	Глобальные
15	Доля предлогов	73,78	Части речи
16	Дисперсия доли глаголов в предложениях	73,74	Части речи
17	Дисперсия доли существительных по предложениям	73,71	Части речи
18	Среднее количество знаков экспрессивной пунктуации	73,58	Редкие
19	Среднее количество существительных в предложении	73,58	Части речи
20	Среднее количество слов, начинающихся с большой буквы	73,57	Глобальные

Также большие различия между русскоязычными и англоязычными текстами видны на редких оборотах. Редкие обороты для русского языка составляют значительную долю наиболее сильных признаков. В то же время, для английского языка соответствующие признаки оказываются гораздо слабее.

Еще одним важным различием с точки зрения классификации является количество полезных признаков. Для русского языка в двадцатке наиболее полезных характеристик F-мера для каждой из них превосходит 73%. Для английского языка только семь наиболее ценных характеристик достигают такого значения. С точки зрения классификации поискового спама данное обстоятельство ухудшает устойчивость алгоритма к попыткам его обойти, так как для английского языка есть меньше надежных признаков.

4.5. Анализ влияния групп признаков

Для того чтобы лучше показать разницу в работе метода для русского и английского языков, мы также сравнили общую ценность всех четырех групп признаков. Для каждой группы признаков был обучен классификатор, использующий только данную группу для обнаружения неестественного текста. Чем лучше работает классификатор, обученный только по группе признаков, тем больший вклад данной группы признаков в классификацию.

Результаты второго эксперимента приведены на рисунке 1. На их основании можно сделать вывод, что для англоязычных текстов метрики разнообразия имеют большее значение для задачи обнаружения неестественных текстов. В то же время, вклад редких характеристик в классификацию оказывается гораздо ниже.

С точки зрения задачи обнаружения поискового спама, сильное влияние одной группы признаков на обучение является негативным фактором, так как позволяет относительно легко снизить эффективность предложенного метода обнаружения неестественных текстов. Спамерам достаточно порождать тексты, в которых распределение Ципфа для слов будет похожим на соответствующее распределение для естественных документов.

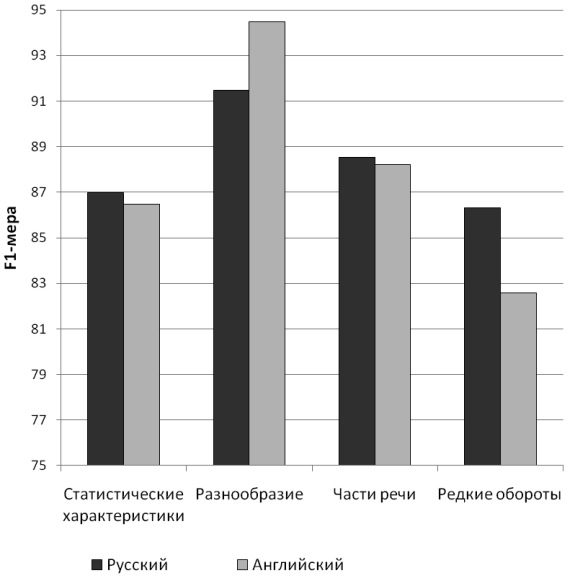


Рис. 1. Вклад групп признаков в классификацию в зависимости от языка

5. Планы

Данное исследование является основой для дальнейшего применения предложенного метода для англоязычных текстов. Мы планируем провести эксперимент по обнаружению поискового спама в коллекции WebspamUK-2007. Данная коллекция специально собрана и размечена для сравнения алгоритмов обнаружения поискового спама.

Для более полного и точного обнаружения спама на данной коллекции потребуется настройка данного алгоритма и на другие виды поискового спама, такие как:

- списки поисковых запросов, вставленные в текст;
- вкрапления ключевых слов в нормальные тексты;
- тексты, составленные из фрагментов из разных источников;
- помимо предложенных в данном методе характеристик также планируется добавление но-

вых признаков, для обнаружения дублированного и украденного содержимого, например, на основе шинглирования [18].

6. Заключение

В данной работе исследовалось применимость алгоритма обнаружения синтезированных текстов, ранее разработанного для русского языка, на англоязычных документах. Было показано, что адаптированный метод позволяет обнаруживать такие тексты с высокой точностью и полнотой.

При этом обнаружено, что вклад различных признаков в работу алгоритма зависит от языка текста. Построенный методом машинного обучения классификатор для английского языка в большей степени полагается на метрики разнообразия текстов, и потенциально менее устойчив к попыткам его обойти.

Разработанный алгоритм также может быть адаптирован для автоматического определения авторства и стиля текстовых документов, а также может быть адаптирован к другим языкам.

Литература

1. *Павлов А. С., Добров Б. В.* Методы обнаружения поискового спама, порожденного с помощью цепей Маркова // Труды 11й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL'2009, Петрозаводск: 2009.
2. *Ntoulas A., Manasse M.* Detecting spam web pages through content analysis // In Proceedings of the World Wide Web conference, ACM Press, 2006. p. 83–92
3. *Piskorski J., Sydow M., Weiss D.* Exploring Linguistic Features for Web Spam Detection: A Preliminary Study // In Proceedings of the 4th international workshop on Adversarial Information Retrieval on the Web, Beijing, China, 2008. p. 25–28.
4. *Гречников Е. А., Гусев Г. Г., Кустарев А. А., Райгородский А. М.* Поиск неестественных текстов // Труды 11й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL'2009, Петрозаводск: 2009.
5. *Mishne G., Carmel D. and Lempel R.* Blocking blog spam with language model disagreement // In Proceedings of the First International Workshop on Adversarial Information Retrieval on the Web, 2005.
6. *Urvoy T., Chauveau E., Filoche P.* Tracking Web Spam with HTML Style Similarities // ACM Transactions on the Web, 2006. Vol. 2, n. 1, Article 3.
7. *Castillo C., Donato D., Murdock V., Silvestri F., Know your neighbors: Web spam detection using the web topology // In Proceedings of SIGIR, ACM, 2007.*
8. *Dubay W. H.* The Principles of Readability // Costa Mesa, CA: Impact Information, 2004.
9. *Фоменко В. П., Фоменко Т. Г.* Авторский инвариант русских литературных текстов // В сб.: Методы количественного анализа текстов нарративных источников. — М.: АН СССР, Ин-т Истории СССР, 1983. с. 86–109.
10. *Парсер* mystem (http://company.yandex.ru/technology/mystem/).
11. *Braslavski P.* Document Style Recognition Using Shallow Statistical Analysis // Proceedings of the ESSLLI 2004 Workshop on Combining Shallow and Deep Processing for NLP, Nancy, France, 2004. p. 1–9.
12. *Quinlan J. R.* C4.5: Programs for Machine Learning // Morgan Kaufmann Publishers, 1993.
13. *Stanford Log-linear Part-Of-Speech Tagger (http://nlp.stanford.edu/software/tagger.shtml).*
14. *Marcus M. P., Marcinkiewicz M. A., Santorini B.* Building a Large Annotated Corpus of English: the Penn Treebank // Computational Linguistics, 1993. Vol. 19 n. 2
15. *Веб коллекция* BY.Web, http://romip.ru/ru/collections/by.web-2007.html.
16. *Yahoo! Research:* "Web Spam Collections". (http://barcelona.research.yahoo.net/webspam/datasets/), Crawled by the Laboratory of Web Algorithmics, University of Milan, (http://law.dsi.unimi.it/).
17. *Gelbukh A., Sidorov G.* Zipf and Heaps Laws' Coefficients Depend on Language // Proceedings of the Second International Conference on Computational Linguistics and Intelligent Text Processing, 2001. p. 332–335.
18. *Зеленков Ю. Г., Сегалович И. В.* Сравнительный анализ методов определения нечетких дубликатов для Web-документов // Труды 9ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL'2007, Переславль: 2007.

К интерпретации видо-временных форм в нарративном режиме: настоящее историческое

Towards interpretaton of tense-aspect forms in narrative register: praesens historicum

Падучева Е. В. (elena708@gmail.com)

Всероссийский институт научной и технической информации

Настоящее историческое трактуется как настоящее нарративное, т. е. как настоящее время нарративного режима. Показано, что настоящее историческое может быть представлено как относительное употребление граммемы времени, когда форма наст. времени соотносится не с моментом речи, а с временным моментом, фиксированным в контексте.

О значении форм грамматического времени мы в Русской грамматике, с. 628 читаем: «Категориальные значения форм времени ориентируются на единую исходную точку — грамматическую точку отсчета. <…> Система форм времени строится на противопоставлении значений одновременности (формы наст. вр.), предшествования (формы прош. вр.) и следования (формы буд. вр.) по отношению к грамматической точке отсчета.»

Обращение к понятию точки отсчета — это шаг назад по сравнению с Грамматикой 1952, в которой говорится (на с. 427): «Категория времени выражает отношение действия к моменту речи или к другому моменту времени, принимаемому за основу временных отношений.»

Формулировка В. В. Виноградова выделяет момент речи в качестве главного ориентира и, соответственно, речевой (иначе — диалогический, дейктический) контекст как главный контекст, в котором интерпретируется временная форма. При этом допускаются и другие ориентиры — очевидно, возникающие в других, вторичных контекстах. Это совпадает с позицией Эстена Даля (Dahl 1985: 25), который настаивает на том, что семантические определения, касающиеся категории времени, должны быть даны прежде всего для основного значения времени, поскольку во вторичных употреблениях форма времени может пересекаться с видовыми формами и теряет четкий смысл.

Чтобы полностью охарактеризовать первичный, т. е. речевой контекст употребления времени, необходимо отграничить его от по крайней мере двух вторичных — нарративного и гипотаксического. Грамматика до последнего времени в упор не видела особенностей употребления формы прош. времени несов. вида в контексте нарратива — и это при том, что примеры, по традиции, приводятся обычно как раз из повествовательных текстов. А между тем в нарративе форма несов. вида прош. времени употребляется в первую очередь в значении одновременности по отношению к ориентиру, возникающему в этом контексте. Таким ориентиром является текущий момент текстового времени. Возьмем фразу:

(А) Бабушка и сейчас *любила* его без памяти.

Фраза (А) однозначно понимается как вырванная из контекста повествования, и потому форма прош. времени интерпретируется в значении одновременности некоторому моменту, «принимаемому за основу временных отношений», — текущему моменту текстового времени, который обозначен словом *сейчас*, а вовсе не значения предшествования этому моменту, как было бы естественно для формы прош. времени; взять хотя бы *Я вас любил*, которое предполагает квази-речевой контекст, так что прош. время выражает предшествование.

Для того, чтобы адекватно описать семантику времени (как и массы других эгоцентрических элементов языка), необходимо обратиться к понятию режима интерпретации (Апресян 1986, 1997, Падучева 1986, 1996), который определяется прежде всего функционированием видо-временных форм. Исходным, первичным, является речевой, или диалогический режим; в нем ориентиром для отсчета времени служит момент речи. А кроме того, есть

нарративный режим, вторичный, и в нем другие ориентиры. Более того, во вторичном режиме иным может быть не только ориентир, но и существо отношения, выражаемого временной формой — в примере (A) форма прош. времени выражает не предшествование, а одновременность.

Понятие режима основано на идеях Э. Бенвениста о повествовательном и речевом дискурсе (Бенвенист 1974: 271). У Бенвениста главным показателем типа текста является именно употребление глагольных форм времени и вида.

Речевой режим (употребления и) интерпретации видо-временных форм предполагает говорящего. Но в нарративе нет полноценного говорящего. Заместителем говорящего оказывается какой-то другой субъект — повествователь (в традиционном нарративе) или (особенно в несобственно прямом дискурсе) персонаж: ориентиром, текущим временем текста, является, так сказать, наст. время одного из них, иначе — время ситуации, в которой они принимают участие.

Режим определяет употребление и интерпретацию видо-временных форм в независимом и в главном предложении. В гипотаксическом контексте возникает еще один ориентир — момент, к которому относится действие, обозначенное глаголом главного предложения; ориентир тут задается синтаксически. Так, в примере (B) словоформа наст. времени *может* выражает одновременность тому моменту, к которому относится клауза *бабушка знала*:

(B) Бабушка знала, что он *может* располагать большими деньгами.

Работа посвящена употреблению формы наст. времени в одном из неречевых режимов — нарративном.

То употребление наст. времени, которое принято называть, следуя античной традиции, настоящим историческим (praesens historicum), я трактую как настоящее нарративное — следуя Грамматике 1952: 484, где это явление рассматривается под именем «настоящее повествовательное». А именно, это употребление формы наст. времени в нарративе с базовым настоящим (в противоположность более обычному нарративу с базовым прошедшим). Пример:

(C) Онегин выстрелил … . Пробили | Часы урочные: поэт | *Роняет* молча пистолет, | На грудь *кладет* тихонько руку | И *падает*. (Пушкин. ЕО)

В Бондарко 1971 показано, что наряду с нарративным употреблением наст. ист. («литературное» наст.ист.; наст.ист. деловой прозы), существует наст. ист. речевого режима, как в примере (D):

(D) *Иду я* вчера по Кузнецкому, вдруг сзади *раздается* свисток.

Эти два употребления наст. ист. различаются по целому ряду параметров. Главным предметом внимания является наст. ист. в нарративном режиме. О речевом употреблении наст. ист. см. раздел 7. Толкование наст.ист. речевого см. в Гловинская 1996.

В существующих трактовках наст. ист. нет упоминания о режиме интерпретации. Между тем, природа наст. ист. существенно прояснится, если сопоставить это употребление формы наст. с другими контекстами, в которых эта форма интерпретируется не в речевом режиме, т. е. не через отношение к говорящему и моменту речи — не как одновременность моменту речи.

Наст. историческое — это текстовое значение; оно невозможно у единичной видовременной формы НСВ наст. Как правило, в предтексте формы, имеющей значение наст. исторического, так или иначе фиксируется некий временной момент, относительно которого локализуется форма наст. ист. Так, в примере (C) этот временной момент задается формой СВ прош. *выстрелил*. Кроме того, наст. историческое может вводить временной момент само по себе, см. пример (1.1) ниже.

Нарративное (повествовательное) употребление имеют не только формы наст., но и формы буд. и прош. времени. Однако с наст. историческим связаны специальные семантические проблемы — оно сложнее, чем прош. нарративное и буд. нарративное.

Во-первых, поскольку в русском языке нет формы СВ в наст.времени, в наст. нарративном не различается СВ и НСВ. Так что специфична у этой видовременной формы не только семантика времени, но и вида — чего нет у прош. и буд. нарративного.

Во-вторых, у формы наст. времени много разных других несобственных значений — не речевых, т. е. не таких, где наст. время означает одновременность с моментом речи, но отличных от нарративного.

Рассмотрю основные аспекты семантики наст. исторического.

1. Наст. историческое и действие в прошлом

Наст. ист. определяется обычно как такое употребление формы наст. времени, которое «повествует о действиях, происходивших в прошлом» (Бондарко 2005: 426). При этом возникают даже две различные формулировки. Одни ученые — в том числе П. С. Кузнецов — считают, что при употреблении наст.ист. говорящий как бы переносится в прошлое. Другие — в том числе В. В. Виноградов — что говорящий переносит прошлое в настоящее, т. е. событие описывается как происходящее в момент речи, см. сопоставление этих формулировок в Бондарко 1971 и Гловинская 1996.

Применительно к «литературному» наст. ист. это противопоставление снимается. Сами категории настоящего и прошлого возникают при соотнесении ситуации с моментом речи. Между тем главное в значении наст. исторического — что оно **не** соотносит ситуацию с моментом речи[1]; и в этом наст. историческое в нарративе подобно прош. и буд. в нарративе, а также всем временным формам в гипотаксическом контексте. «Литературное» наст. ист. может быть представлено примером (1.1):

(1.1) По базарной площади *идет* полицейский надзиратель Очумелов (Чехов. Хамелеон).

Про «литературное» наст. ист. в Бондарко 1971: 146 говорится: «здесь “я” отсутствует. <…> Нет живой связи, актуального соотношения ни с “я”, ни с моментом речи». (Имеется в виду, отсутствует “я” как говорящий — в повествовании о прошлом “я” уже не говорящий.)

Подробный обзор трактовок наст.исторического дан в Dickey 2000: 126–154. Преобладает точка зрения, что наст.ист. сдвигает дейктический центр (т. е. момент речи) в прошлое — во время события[2]. Общепризнано, однако, что заместителем говорящего в нарративе является повествователь, и в нарративе с базовым прошедшим повествователь тоже «видит» события в синхронной перспективе. В чем же тогда отличие наст. ист. от прош. нарративного, как в примере (1.2)?

(1.2) По базарной площади *шел* полицейский надзиратель Очумелов.

Для описания различия между наст. и прош. нарративным следует принять во внимание слушающего (читателя, адресата): прош. нарративное устанавливает временную и пространственную дистанцию между ситуацией и адресатом, а наст.ист. приглашает адресата наблюдать ситуацию изблизи (см. Бондарко 1971: 143, Падучева 1996: 287–290; Гловинская 1996: 456).

Ю. С. Маслов (в Лингвистическом энциклопедическом словаре) пишет: «формы наст.ист. повествуют о событиях, имевших место в прошлом, но изображаемых как современные наст. моменту». Но тогда в приводимом им примере —

(1.3) В 1725 году Петр I *умирает* — нужно признать, что 1725 год изображается как современный наст. моменту — что, очевидно, не так.

2. Наст. историческое как относительное время

Наст.ист. (нарративное) можно трактовать как относительное употребление формы времени: форма наст. в этом употреблении соотносит ситуацию не с моментом речи, как в исходном, речевом значении, а с некоторым моментом, фиксированным в контексте. При интродуктивном употреблении наст.ист., как в примере (1.1), сама форма наст.ист. задает тот временной момент, который используется для последующих отсылок. Что касается «цепочки» форм наст.ист., то в этом контексте временная отнесенность определяется видовым значением формы — процессным или событийным (см. разделы 5 и 6); процессное значение дает одновременность, событийное — последовательность.

Ю. С. Маслов считает относительным употреблением только гипотаксическое, а наст. ист. относит к переносным. Между тем в статье Шмелев 1960 предлагается трактовка наст. исторического как относительного употребления формы наст. Д. Н. Шмелев приводит примеры типа *И вот началась вьюга <…> Вокруг все кипит точно в котле* (Горький. Варенька Олесова), говоря, что «относительное значение формы, т. е. прикрепленность к объективному прошедшему, опирается на ее сочетание с формами грамматического прошедшего» — в самом деле, наст. *кипит* соотносится с прош. *началась*.

Что же касается наст. ист. речевого (в интродуктивном употреблении), то ему свойственна м е т а ф о р и ч н о с т ь, и в этом смысле его можно считать переносным, см. подробнее в разделе 7.

В Бондарко 1971: 227 приводится пример нарратива, где форма наст. времени относит ситуацию к плану будущего (употребление квалифицируется как «настоящее при обозначении будущих действий»):

(2.1) Она старомодно мечтала о том, как он <…> *появится* перед окном домика верхом на лошади <…> Они *войдут* в гостиную, оба взволнованные … Вдруг он *берет* ее за руки выше локтей, *привлекает* к себе … (А.Толстой. Хождение по мукам)

Ясно, что употребление форм *берет* и *привлекает* нарративное, и, как отмечает А. В. Бондарко, у этого употребления много общего с наст. историческим. Однако есть разница: форма наст. в значении будущего действия не может быть интродуктивной.

3. Наст. историческое как текстовое употребление формы наст.

Итак, наст. ист. — это такое употребление формы наст. времени, которое соотносит данную

[1] О том, что к повествовательному тексту неприменимо понятие момента речи, см., в частности, Плунгян 2000: 266.

[2] Возникает парадоксальная формулировка — форма настоящего времени означает, что «говорящий “как бы” переносится в прошлое».

ситуацию с некоторым моментом, фиксированным в тексте, см. пример (С). В специальном контексте (который в Бондарко 1971 выделяется как «биографическое настоящее») момент может фиксироваться датой, как в (1.3). С другой стороны, форма наст. ист. фиксирует момент для последующей отсылки. В любом случае, наст. ист. — это текстовое время; оно невозможно в изолированном высказывании:

(3.1) а. Я *иду* по Кузнецкому [наст. речевое: так можно сказать по мобильному телефону]
б. Иван *идет* по Кузнецкому [наст. речевое: говорящий выступает в роли наблюдателя за Иваном — тоже в речевом режиме]

(3.2) Иван *идет* по Кузнецкому. <Вдруг сзади *раздается* свисток.> [наст.ист.]

В (3.1а) и (3.1б) форма наст. — это наст. речевое, а в (3.2) та же фраза, но в контексте следующей, воспринимается как интродукция момента времени, так что в (3.2) наст. время глагола в той же фразе (3.1б) понимается как наст. ист. (так при именной анафоре именная группа становится антецедентом, когда есть заместитель).

Предложение (3.3) допускает две интерпретации, в зависимости от правого контекста — в а) наст. предстоящее, в б) — наст. ист.:

(3.3) В Москве *открывается* выставка Пикассо;
а) В Москве *открывается* выставка Пикассо. Я обязательно пойду.
б) В Москве *открывается* выставка Пикассо. На нее устремляются толпы людей.

Предложение, обеспечивающее интродукцию временного момента, должно иметь определенную коммуникативную структуру. Так, в (3.4) (пример из Падучева 1996) первая фраза в (3.4а) может быть интродукцией текущего момента текста, а фраза (3.4б) — не может:

(3.4) а. Привезли меня в больницу. *Лежу* на койке.
б. А в больницу меня *привезли* с брюшным тифом. (Зощенко)

4. Способы интродукции момента времени

Какие могут быть способы интродукции момента времени, на который ориентирована форма наст. ист.?

- Самый естественный способ интродукции текстового времени — это форма СВ:

(4.1) Она сидела вязала. Вдруг бросила вязать и *смотрит* на меня.

(4.2) Федоров <…> запер его [голубя] в сарай <…>. Но вдруг об этом деле *узнает* председатель сельсовета Егоров. Он моментально *бежит* к этому жителю и ему *говорит* … (Зощенко. Усердие не по разуму)

(4.3) Зашел тут ко мне приятель. *Рассказывает* такую историю.

- Момент текущего (текстового) времени, необходимый для наст. ист., может быть задан обстоятельством времени:

(4.4) а. В 1994 году Солженицын *приезжает* в Москву.
б. После полудня к хозяину *приезжает* очень высокий и очень толстый мужик <…> «Вы из России?» — спрашивает он меня. (Чехов. Из Сибири; пример из Бондарко 2005: 334)

При интродуктивном употреблении в отсутствие показателя времени, как в (3.2), момент времени задается самой формой наст.ист. Параметры контекста, которые заставляют понять форму наст. времени как обозначающую «действие в прошлом» подлежат дальнейшему изучению. В самом общем виде можно сказать, что человек интерпретирует форму наст. как относящую ситуацию к прошлому ввиду того, что, по тем или иным, быть может, прагматическим, причинам не может понять ее в наст. речевом (например, когда человек открывает дверь и говорит: *Иду я по Кузнецкому* …).

5. Наст. нарративное в сравнении с прош. нарративным

Контекст позволяет различить у формы наст. ист. два видовых значения: процессное — как бы значение несов. вида, см. (3.2), и событийное — как бы значение сов. вида, см. (4.4).

Сопоставим наст. ист. (т. е. настоящее нарративное) в процессном значении с несов. видом прошедшего нарративного. Между этими формами имеются следующие сочетаемостные различия, вытекающие из семантики времени:

- Наст. историческое сочетается с *вот*, а прош. нарративное — нет:

(5.1) Вот *едет* могучий Олег со двора (Пушкин); *Вот *выезжал* Олег из двора;

(5.2) И вот *стою* на площадке (Зощенко); * И вот *стоял* на площадке.

Это различие можно объяснить оппозицией «ближнее–дальнее»: повествователь в любом слу-

чае — и при наст. и при прош. времени является синхронным наблюдателем. Различие в том, что форма наст. как бы приглашает адресата к участию в ситуации, создавая видимость общего поля зрения, а прош. отдаляет адресата от места действия.

Пример (5.3) мог бы считаться опровергающим это положение:

(5.3) Вот, тяжело ступая по коридору, дневальные *несли* одну из восьмиведерных параш. (Солженицын. Один день Ивана Денисовича)

Но такое употребление порождено свойственным этой повести свободным косвенным дискурсом, в котором нарушаются многие нормы традиционного нарратива. Замена *несли* на *несут* дала бы более нормативный текст.

- Наст. историческое свободно употребляется для обозначения действия, наступающего после того, которое обозначено глаголом СВ, тогда как НСВ прош. может быть в этом контексте неуместно:

(5.4) а. Он схватил меня за руку и *тянет*;
б. *Он схватил меня за руку и *тянул*.

Вместо (5.4б) надо сказать:

(5.5) Он схватил меня за руку и *стал тянуть* (или *потянул*).

Аналогичный пример:

(5.6) а. *Снимает* шапку и *крестится*;
б. **Снял* шапку и *крестился*.

Это существенное различие между наст. и прош. нарративным можно объяснить тем, что в контексте наст. ист. форма несов. вида приобретает значение начинательности, которое отсутствует у НСВ прош. в русском языке: форма НСВ прош. выделяет срединную фазу действия. Отсутствием начинательности у русского НСВ в прош. времени можно объяснить различия в употреблении формы НСВ в русском и чешском (ср. Петрухина 2000: 84), в котором такие сочетания видовых форм, как в (5.4б), (5.6б), вполне нормативны; примеры (5.4б), (5.6б) — буквальные переводы чешских примеров Е. В. Петрухиной. В Dickey 2000 эта конструкция названа “чешский контекстно-обусловленный имперфектив прошедшего времени”.

В Зализняк, Шмелев 2000: 28 говорится: «При повествовании в наст. историческом говорящий заменяет формы прошедшего формами настоящего». Примеры (5.4), (5.6) показывают, что это не совсем так: текст в «исходной» форме (так сказать, до замены прош. на наст.) может не существовать.

6. Наст. историческое и совершенный вид

Известны сочетаемостные свидетельства в пользу того, что у наст. исторического есть событийное значение — как бы значение сов. вида.

- Цепочка глаголов в наст. ист. событийном обозначает последовательные события, что свойственно глаголам СВ:

(6.1) … *выступает* холодный пот. Он *кладет* ножницы и *начинает* тереть себе кулаком нос.

В примере (С) форма *роняет* обозначает событие, следующее за *выстрелил* (и *пробили*).

- Наст. ист. сочетается с наречиями типа *неожиданно*, *внезапно, моментально, вдруг*, с которыми может сочетаться только сов., а не несов. вид; так, возможно (6.2а) и (6.2б), но не (6.2в):

(6.2) а. Неожиданно из-за отдаленного кустарника *выползает* луна [наст. ист.]
б. Неожиданно из-за отдаленного кустарника *выползла* луна [СВ];
в. *Неожиданно из-за отдаленного кустарника *выползала* луна [несов.прош.]

- Аспектуальное сходство наст. ист. событийного и СВ обнаруживают реверсивные глаголы. Так, в (6.3а), с наст. ист., наблюдатель остается в той временной точке, когда я еще нахожусь у приятеля, как и при СВ; а реверсивный глагол в прош. речевом в несов. виде возвращает наблюдателя в исходное положение, см. (6.3в):

(6.3) а. *Захожу* я к приятелю [в момент наблюдения я у приятеля];
б. Я *зашел* к приятелю [в одном из пониманий — в момент наблюдения я у приятеля];
в. Я *заходил* к приятелю [и ушел].

Однако между наст. ист. и СВ есть и различия. Примечательно, например, что наст. ист. невозможно в контексте *только что*, хотя глагол СВ в этом контексте абсолютно нормален (NB вышеупомянутую начинательность наст.исторического):

(6.4) Только что *кончился* (**кончается*) математический съезд.

Понимание наст. ист. в событийном значении невозможно также в контексте *уже*:

(6.5) Уже *кончается* математический съезд = ‘вот-вот кончится’, а не ‘уже кончился’.

7. Наст. историческое и обстоятельства времени

Примеры употребления наст. ист. в контексте обстоятельств времени с дейктическим значением[3]:

(7.1) а. *Иду* вчера по Гостиному двору, — вдруг он! [Д. В. Григорович. Недолгое счастье (1884)];
б. *Поднимаюсь* недавно по лестнице в гостинице «Красная». [Роман Карцев. «Малой, Сухой и Писатель» (2000–2001)];
в. Много лет назад *появляются* первые механические часы, и с тех пор идет постоянное состязание в усовершенствовании прибора (пример из Гловинская 1996).

Слова *вчера* и *недавно* маркируют речевой режим интерпретации предложения. Неудивительно поэтому, что в примерах часто встречается субъект 1 лица:

(7.2) Ложимся вчера спать, *я спрашиваю*: — Вова, а сколько рекорд? [Л. К. Бронтман. Дневники и письма (1932–1942)]; *Прихожу* вчера в контору за справкой. Швейцар отворяет дверь: [П. П. Гнедич. Книга жизни (1918)]; *Получаю* вчера письмо ваше, читаю и недоумеваю. [Ф. М. Достоевский. Роман в девяти письмах (1847)]; *Приезжаю* недавно в отпуск, а она замужем. [Л. К. Бронтман. Дневники и письма (1943–1946)]

Интродуктивная структура предложения делает слова *вчера* и *недавно* атоническими, и они, как энклитики, по закону Вакернагеля попадают в позицию после первого ударного слова во фразе. Примеры (7.3а, б) показывают, однако, что субъект 3 лица тоже возможен, но тогда подлежащее 3 лица отделяется от сказуемого тем, что в Зализняк 2008 называется ритмико-синтаксическим барьером:

(7.3) а. Наш сосед радиотехник Лёнька Шалыт *заходит* вчера после дежурства вздремнуть. Батюшки мои, в жилище полметра воды! [Василий Песков. Белые сны (1964)];
б. … а мясник *говорит* вчера жене: вижу я, говорит, что вы с вашим мужем не больше, как жулики. [Л. Н. Андреев. Москва. Мелочи жизни (1901–1902)].

Пример (7.4) производит комическое впечатление, потому что слово *вчера* — первичный эгоцентрик и уместно в речевом контексте, а (7.4) напоминает зачин анекдота, т. е. типичного нарратива.

(7.4) ?*Встречается* вчера президент с премьер-министром (пример из Всеволодова, Ким Тэ Чжин 2002).

Если адвербиала в предложении нет, то по умолчанию подразумевается ‘недавно’ (противное должно быть оговорено):

(7.5) Прихожу я в контору за справкой.

При этом отстояние ситуации от наст. момента должно быть соразмерно с важностью, примечательностью ситуации, ср. нормальное (7.6а) и странное (7.6б):

(7.6) а. Приезжаю *я в прошлом году* в Нью-Йорк;
б. ?Иду *я в прошлом году* по Кузнецкому.

Итак, в предложениях с обстоятельством времени наст. ист. выражает синхронность моменту, который задается обстоятельством.

В Гловинская 1996: 454 в толкование наст. ист. в речевом контексте включается компонент ‘Говорящий мыслит себя в прошлом, и действие как бы происходит на его глазах’. Как, однако, в этом случае интерпретировать предложение (7.1а) *Иду вчера по Гостиному двору — вдруг он*? Наречие *вчера* определенно предполагает говорящего, который мыслит себя сегодня, т. е. именно в настоящем, а не в прошлом, к которому относится ситуация. Выход из этого противоречия мыслится в том, что перенос момента речи в прошлое является метафорическим, NB *как бы*.

Можно думать, за эффект наглядности, который в (7.1а) несомненно присутствует, отвечает не говорящий, а тот наблюдатель, который входит в семантику несов. вида (в данном случае имеющего процессное значение). Говорящий как субъект временного дейксиса находится в настоящем, а как субъект наблюдения — в том прошлом, к которому относится ситуация. В разговорном наст. ист. категориальный диссонанс между действием в прошлом и формой наст. времени создает экспрессию (см. Бондарко 1971: 145) — то, что А. В. Бондарко называет «образной актуализацией прошлого». И носителем этой экспрессии не может быть никто, кроме говорящего. Так что наст.ист. речевого режима — это метафорическое настоящее.

Как справедливо указывает А. В. Бондарко, идея наглядности проходит отнюдь не для всех употреблений наст.исторического[4]. Например, в ее нет в (7.7):

(7.7) К концу XII века колесные часы уже существовали. В 1232 году Данте *упоминает* о колес-

[3] Датированные примеры — из Национального корпуса русского языка.

[4] См. обсуждение дискурсивных эффектов наст. исторического в Dickey 2000: 127.

ных часах с боем. В 1288 году *устанавливают* башенные часы в Westminster Hall в Лондоне. С XIV в. башенные колесные часы *появляются* в различных городах. [Энциклопедия Брокгауз и Эфрон]

Итак, мы можем заключить, что наст. ист. (повествовательное), будучи текстовым временем, соотносит действие с моментом, фиксированным в тексте, а в интродуктивном употреблении (при отсутствии обстоятельства времени) само вводит этот момент. Строго говоря, отнесенность действия к прошлому формой наст. ист. не выражается, см. пример (3.3): когда мы понимаем, что действие происходит в прошлом, мы осознаем, что форма наст. выражает наст. историческое, а не наоборот.

В прошлое перенесен не момент речи, а момент наблюдения. Момент речи в интерпретации формы наст. исторического не участвует. Участвует говорящий-повествователь как субъект экспрессии и как контрагент адресата[5].

[5] Автор благодарит Ю.Д.Апресяна за поправки и критические замечания, высказанные по поводу первоначального варианта работы.

Литература

1. *Апресян Ю. Д.* Лексическая семантика: Синонимические средства языка. М.: Наука, 1974.
2. *Апресян Ю. Д.* Дейксис в лексике и грамматике и наивная модель мира // Семиотика и информатика. Вып. 28. М.: 1986. С. 5–33.
3. *Апресян Ю. Д.* Лингвистическая терминология Словаря // НОСС 1997. С. XVI–XXXIV.
4. *Бенвенист Э.* Общая лингвистика. М.: Прогресс, 1974.
5. *Бондарко А. В.* Вид и время русского глагола. М.: 1971.
6. *Бондарко А. В.* Теория морфологических категорий и аспектологические исследования. М.: 2005.
7. *Всеволодова М. В., Ким Тэ Чжин.* Система значений и употреблений форм настоящего времени русского глагола (в зеркале корейского языка). М.: 2002.
8. *Гловинская М. Я.* Две загадки praesens historicum // Русистика. Славистика. Индоевропеистика. Сб. статей к 60-летию А.А.Зализняка. М.: Индрик, 1996. С. 451–458.
9. *Грамматика* 1952 — Грамматика современного русского литературного языка: Т. I. Фонетика и морфология. М.: Наука, 1952.
10. *Зализняк А. А.* Древнерусские энклитики. М.: Языки славянских культур. М.: 2008
11. *Падучева Е. В.* Семантика вида и точка отсчета // Изв. АН СССР. Сер. лит. и яз. 1986. Т. 45. № 5. С. 413–424.
12. *Падучева Е. В.* Семантические исследования: Семантика времени и вида в русском языке. Семантика нарратива. М.: Языки рус. культуры, 1996.
13. *Петрухина* 2000 — Петрухина Е. В. Аспектуальные категории глагола в русском языке в сопоставлении с чешским, словацким, польским и болгарским языками.М.: МГУ, 2000.
14. *Русская* грамматика. Т. 1–2 / Отв. ред. Н. Ю. Шведова. М.: 1980.
15. *Шмелев Д. Н.* Абсолютное и относительное употребление форм времени русского глагола. Русский язык в национальной школе. 1960, № 6. С. 3–10.
16. *Dahl Ő.* Tense and Aspect Systems. Oxford-N.Y.: Basil Blackwell, 1985.
17. *Dickey S. V.* Parameters of Slavic aspect. A cognitive approach. CSLI publications, Stanford, California, 2000.

Самоисправления говорящего в японском устном нарративе: анализ корпусных данных

Self-repairs in japanese narrative discourse: a corpus-based case-study

Подлесская В. И. (podlesskaya@ocrus.ru),
Комарова А. Д. (komarovochka@gmail.com)

Российский государственный гуманитарный университет (РГГУ)

Рассматриваются типы самоисправлений говорящего в неподготовленном устном японском дискурсе. На основе сопоставления с данными русского языка выявляются лингвоспецифические и универсальные механизмы самоисправлений.

1. Постановка вопроса

В работе предпринято корпусное исследование самоисправлений говорящего в японской неподготовленной монологической речи.1Материалом послужил корпус устных рассказов по картинкам и их последующих пересказов по памяти (23 рассказа и 23 пересказа), общий объем корпуса — 4545 слов, общая продолжительность звучания — 44,25 мин.

Транскрибирование и разметка речевых сбоев производилась по методике предложенной в Подлесская, Кибрик 2009. В этой же работе были предложены и апробированы на материале русских устных рассказов базовые критерии классификации самоисправлений. На эти критерии мы и опирались при работе с японским материалом. Как мы покажем ниже, этот набор критериев, в целом, продемонстрировал универсальную применимость к языку иного грамматического строя, однако конкретные механизмы самоисправлений, а также количественное распределение отдельных типов самоисправлений в японском языке обнаружили весьма поучительную специфику.

2. Коррекции vs. редактирование

Прежде всего, вслед за вышеупомянутой работой Подлесская, Кибрик 2009, мы выделили два основных режима самоисправления: — (он-лайн) коррекцию, или собственно **коррекцию**, и ретроспективную коррекцию, или **редактирование**.

В режиме (он-лайн) коррекции говорящий реагирует на обнаруженную проблему немедленно, поэтому при коррекциях речевой отрезок до точки прерывания обладает незавершенностью по совокупности лексических, грамматических и просодических критериев. Очень часто в точке прерывания имеется обрыв слова, но даже там, где обрыва слова нет, имеются симптомы синтаксической и дискурсивной неполноты текущего отрезка. После прерывания говорящий заменяет забракованный фрагмент на другой или повторяет первоначально забракованный фрагмент и продолжает развертывать речь таким образом, чтобы забракованный фрагмент мог быть безболезненно «стерт» и материал до забракованного фрагмента и материал после него, сомкнувшись, образовали бы правильную, т. е. удовлетворяющую говорящего последовательность. Так, в примере (1):

(5) YUK T1:30[2]

30.	otokonoko	ni	..(0.2)	chiisa	na		
	мальчик	DAT		маленький	ATR		
	kuruma ||	omochya	no	kuruma	o	kat-te	
	машина	игрушка	GEN	машина	ACC	купить-CNV	

Мальчику маленькую машинку …. игрушечную машинку купив…

[1] Исследование выполнено при поддержке РГНФ, грант 09-04-00106а

[2] После номера примера приводится его индекс в корпусе, включающий указание на имя говорящего, код рассказа и номер строки.

имеется забракованный фрагмент *chiisa na kuruma* ‘маленькую машину’, который после точки прерывания (отмеченной знаком ||) заменен на откорректированный фрагмент *omochya no kuruma* ‘игрушечную машину’. Удаление забракованного фрагмента дает текст, грамматически приемлемый и ситуационно уместный с точки зрения говорящего.

Ретроспективная коррекция, или редактирование, это менее «авральный» режим преодоления обнаруженной проблемы. Он состоит в том, чтобы отредактировать уже готовый и транслированный фрагмент дискурса постфактум, т. е. уже по завершении проблемного отрезка информировать слушающего о том, что этот отрезок подлежит уточнению или исправлению. Как иллюстрирует пример (2), при редактировании могут (но не обязательно должны!) использоваться особые дискурсивные маркеры, сигнализирующие о преодолении затруднения. Типичным примером является японское междометие «неожиданной находки» *A!* (произносится на сильном выдохе с резким нисходящим тоном, ср. близкий и по форме и по функции русский речевой жест). Эпизод редактирования может сопровождаться эксплицированным признанием ошибки (*Mactigatta!* ‘[Это было] неправильно!’) и уточняющими выражениями со значением ‘точнее’, ‘то есть’ и т. п., часто — совместно со стандартными способами выражения отрицания и противопоставления:

(6) YAM R1:10–13

10. ….(1.5) kono keshoohin wa zenbu
этот косметика TOP[3] все

….(1.1) roshia de tukut-ta
Россия INS сделать-PST

Эта косметика вся изготовлена в России.

11. ..(0.1) A-a machigat-ta
а ошибиться-PST

А, нет! (Букв.[я] ошибся)

12. ..(0.3) Chu= || chuugokusei de ==
кита= китайский.товар COP.CNV

Кита= … китайские товары явл= …

13. Kono subete chugokusei da.
этот все китайский.товар COP.PRS

Это все китайские товары.

Использование режима редактирования в японском языке может приводить к возникновению постпозитивных уточнений, нарушающих нормативный японский синтаксис, характеризующийся последовательным левым ветвлением и заключительной позицией сказуемого. Так, в примере (3) имеет место так называемая правая дислокация — вынос в позицию после заключительного сказуемого элементов, синтаксически от него зависящих; в данном случае вынесена именная группа *asagohan ni* ‘на завтрак’, которая при «идеальном речепорождении», т. е. при отсутствии речевого сбоя, должна была бы располагаться в абсолютном начале клаузы, но была «забыта» говорящим, а затем восстановлена пост-фактум:

(7) KIM T2: 3–4

3. …(0.5) Sonoato ….(1.0) mmm(0.6) ..(0.4)
Потом

koocha to ..(0.3) pizza o tabe-mash-ita,
черный.чай и пицца ACC есть-ADR-PST

После этого он съел пиццу с чаем,

4. asagohan ni.
завтрак DAT

на завтрак.

Интересно, что планирование выноса вправо говорящий осуществляет заблаговременно: перед дислоцированным элементом часто не бывает паузы, а предшествующее ему сказуемое может интонироваться как незаключительное в высказывании, в частности, падение частоты основного тона может быть не в самый низкий уровень, типично характеризующий конец высказывания у конкретного говорящего.

Дислоцированные уточнения могут внедряться и внутрь клаузы. В примерах (4) и (5) говорящий дважды — сначала в рассказе по картинке, а потом при пересказе того же сюжета по памяти — после именной группы *kodomotachi* ‘дети’ использует вставку, уточняющую «состав семьи», причем в (4) при выходе из вставки для обозначения возвращения к основной линии изложения используется специальный маркер, восходящий к противительному союзу со значением ‘однако’:

(8) WAK T1: 10–13

10. sokonohito ga **..(0.2)** jibun no kodomo-tachi –
мужчина TOP сам GEN ребенок-PL

Тогда мужчина со своими детьми

11. …(0.5) ee(0.1) musume to musuko des-u
дочь и сын COP.ADR-PRS

keredomo
однако

— а это были дочь и сын —

12. **…(0.6)** – ittai nani o kat-tara
наконец что ACC купить-COND

---

[3] В морфологической строке примеров используются следующие сокращения: ACC — аккузатив, ADR — адрессив, ATR — атрибутивная форма, AUX — вспомогательный глагол, AUX.S — вспомогательный глагол окончательного действия *shimau,* CNV — деепричастие, COND — условная форма, COP — связка, DAT — датив, EVD — эвиденциальность, GEN — генитив, HON — гоноратив, INS — инструменталис, NEG — отрицание, NML — субстантиватор, NOM — номинатив, PL — множественное число, PMT — презумптив, PRG — прогрессив, PRS — настояще-будущее время, PST — прошедшее время, Q — вопросительная частица, QUOT — цитационный союз, TOP — топик, VRB — вербализатор.

i-i da-roo ka
хороший-PRS COP-PMT Q

в конце концов, что же было бы хорошо купить,

13. **..(0.2)** to soodan shi-mash-ita
QUOT совет VRB-ADR-PST

посоветовался.

(9) WAK R1: 9–12

9. **....(2.1)** soshite **..(0.2)** otoko wa **..(0.2)**
тогда мужчина TOP

jibun no kodomo-tachi –
сам GEN ребенок-PL

Мужчина своим детям

10. ...(0.5) ee(0.2) musume to ..(0.2) musuko ni
дочь и сын DAT

— дочери и сыну —

11. **...(0.7)** ittai o-kaa-san ni nani o
наконец мама DAT что ACC

age-tara i-i da-roo ka
подарить-COND хороший-PRS COP-PMT Q

«В конце концов, что было бы хорошо подарить маме?»

12. to i-imash-ita
QUOT сказать-ADR-PST

сказал.

Фактически, в примерах типа (3)–(5) мы наблюдаем, как потребность в самоисправлении оказывается более мощным фактором речепорождения, чем структурные правила. Компромиссом между нормативным синтаксисом, запрещающим расположение зависимых элементов справа от вершины, и дискурсивно обусловленной необходимостью самоисправления оказываются конструкции, в которых дислоцированный элемент снабжается глаголом-связкой. Так, в следующем примере, в результате редактирования придаточное цели дислоцируется вправо от главной клаузы — вместо стандартного для японского языка порядка «придаточное-главное» («маме ко дню рождения подарок купить чтобы, отец по магазинам пошел»), возникает порядок «главное-придаточное» («отец по магазинам пошел, маме ко дню рождения подарок купить чтобы»). При этом за придаточным непосредственно следует связка, букв. «отец по магазинам пошел, маме ко дню рождения подарок купить чтобы [это] есть» (о правой дислокации придаточных в японском языке см. Ohori 1996) . Дислоцированные элементы со связкой по функции близки постпозитивным обстоятельственным клаузам с *это* в русской разговорной речи (*это чтобы ..., это потому что...* и т. п.):

(10) KIM R1:1–2

1. Aru hareta hi no gogo otoo san wa
Один ясный день GEN после.обеда отец сан TOP

kaimono e dekake-mash-ita.
покупки в выйти-ADR-PST

В один ясный день отец пошел по магазинам.

2. ....(1.5) Okaa san e no tanjoobi purezento
Мама сан в GEN день.рождения подарок

o ka-u tame des-u.
ACC гупить-PRS чтобы COP.ADR-PRS

[Это] чтобы маме ко дню рождения подарок купить.

Как показывают наши количественные данные, самоисправления в режиме редактирования встречаются значительно реже, чем собственно коррекции (14 % против 86 % , о количественном распределении самоисправлений см.подробнее ниже)

3. Основные типы коррекций

Для классификации он-лайн коррекций используются следующие параметры (подробнее о них, см. Подлесская, Кибрик 2009):

1. Затрагивает ли коррекция структуру дискурса (макрокоррекции) или только один минимальный дискурсивный шаг, типично — клаузу (микрокоррекции)?
2. Каков объем забракованного фрагмента?
3. Какая операция осуществляется с забракованным фрагментом — он повторяется, частично изменяется или полностью отменяется?
4. Расположены ли забракованный фрагмент и его откорректированный коррелят контактно или дистантно?

Абсолютное большинство (93 %) коррекций в корпусе являются микрокоррекциями, т. е. сбоями, не выходящими за пределы клаузы, ср. пример (1) выше. Макрокоррекции, затрагивающие более одного узла дискурсивной структуры, могут быть проиллюстрированы примером (7), в котором сбой приводит к замене конструкции с инфинитной формой на конструкцию с причинным союзом и финитным глаголом:

(11) DAI R2:7–11

7. ...(0.7) eee(0.8) ....(1.2) Keredomo ..(0.1) kare
но он

wa ..(0.4) ee(0.3) suko= ||
TOP немн =

sukoshi yopparat-te shimat-te ==
немного пьянеть-CNV AUX.S-CNV

Но, немного опьянев ...

8. ....(8.2) Shika==
Однак=

Однак= ....

9. ee(0.4) yopparat-te shimat-ta node
пьянеть-CNV AUX.S-PST так.как

Поскольку опьянел,

10. ..(0.2) ee(0.4) ….(1.4) sukii o shi-te
лыжи ACC делать-CNV

i-te
AUX.PRG-CNV

катаясь на лыжах,

11. ..(0.1) ee(0.2) koron-de shimai-mash-ita
падать-CNV AUX.S-ADR-PST

упал.

По второму критерию — объему забракованного фрагмента — коррекции в корпусе делятся на два практически равных класса: в половине случаев самоисправлению подвергается мелкий фрагмент (к этому классу мы относим служебные слова, оборванные фрагменты полнозначных слов и их комбинации), другая половина коррекций затрагивает крупные фрагменты (полнозначные слова и их комбинации). В примерах (1) и (7) были приведены коррекции крупных фрагментов: в (1) это именная группа (существительное с зависимым определением), в (7) — это аналитическая глагольная форма. Приведем пример коррекции мелкого фрагмента — в (8) забракована часть полнозначного слова *sut=* (неудавшаяся попытка прознести искомое слово *sutando* ‘лампа’):

(12) YUK T1:4

4. ...(0.8) De hajime wa-aa ..(0.5) baggu ..(0.2)
И сначала TOP сумка

ya-a ..(0.4) mmm(0.5) sut= || ...(0.6) sutando
и лам= лампа

o mi-te i-ru ka-a
ACC смотреть-CNV AUX.PRG-PRS Q

Сначала (он) что-ли разглядывает сумки, лам= … лампы…

Наибольшая вариативность коррекций наблюдается по третьему критерию — по типу операции, которой подвергается забракованный фрагмент. В более чем в половине случаев (53.9 %) происходит повтор забракованного фрагмента, ср. (8). Повтор используется как средство для того, чтобы выиграть время для планирования текущего или последующего отрезка дискурса. Если повтор связан с планированием текущего отрезка, то автомониторинг дает следующий результат: произнеся некоторый фрагмент говорящий прерывается, ощутив сомнение в том, что этот фрагмент соответствует изначальной программе, затем отвергает эти сомнения и повторяет данный фрагмент, сигнализируя о том, что вербализации более удачной, чем первоначально предложенная, он не подобрал.

Модификации — замены забракованного фрагмента с полным или частичным сохранением его значения — составляют 31,6 % от общего числа коррекций. Типичным примером модификации является (1), где говорящий заменяет указание на размер, ‘маленькую’, на более точное определение «размер плюс функция» — ‘игрушечную’. Заслуживает внимания тот факт, что почти половину всех модификаций составляют случаи, затрагивающие не лексическое наполнение высказывания (т. е.замена слова на более уместное), а его грамматическую форму. Так, в примерах (9) и (10) говорящий в ходе самоисправления меняет грамматическую форму глагола. В (9) фрагмент *ka-i=* представляет собой оборванную форму из парадигмы презенса глагола со значением ‘купить’ (она могла бы быть, например, началом вежливой формы непрошедшего времени индикатива или формы желательного наклонения), в ходе самоисправления говорящий меняет «грамматический дизайн» и выбирает отрицательную форму потенциалиса *ka-e-nai* ‘не может купить’:

(13) MAI R1:9–10

9. kai-i= ==
купи=

10. ka-e-nai node
покупать-POT-NEG потому.что

Так как (он) покупае=… не может купить…

Сходным образом в (10) говорящий забраковал основу глагола *nom=* ‘купить’, которая используется с формами группы презенса — еще до того, как выбрать окончание. Вместо этого, он предпочел форму прошедшего времени *non-da* ‘выпил’, для которой используется основа *non-* с чередованием финали основы *m/n*:

(14) DAI T2:4

4. minna de ….(2.4) tanoshiku ….(2.7) o-sake
вместе INS весело HON-сакэ

o nom= || …(0.4) non-da n ..(0.3)
ACC пьё= пить.PST NML

des-u keredomo
COP.ADR-PRS хотя

И хотя (он) вместе со всеми весело пьё= … пил сакэ…

Со сменой глагольной формы связано и самоисправление в (11), где забраковано наречие степени *totemo* ‘очень’, которое в японском языке допускается только с положительными формами глагола. Как выясняется, говорящий предпочел отрицательную форму; эта отрицательная форма появляется, как и положено, только в абсолютном конце клаузы, однако уже в точке выбора наречия степени наречие *totemo*

исправляется на *amari* — наречие степени, которое используется только с отрицательными формами глагола. Следовательно, уже в этой точке речепорождения первоначально задумывавшаяся положительная форма глагола была перепланирована и заменена на впоследствии и реализованную отрицательную:

(15) KEI R1:20

20. ...(0.6) Okaasan wa ..(0.4) totemo || ...(0.9)
мама TOP очень

mmm(0.5) amari ..(0.3) yorokon-de i-na-i
совсем радоваться-CNV AUX.PRG-NEG-PRS

yoo desh-ita.
EVD COP.ADR-PST

Похоже, мама очень ... совсем не обрадовалась.

«Грамматические» сбои могут быть связаны не только со словоизменением, но и с синтаксисом — в частности, с синтаксической упаковкой клаузы. Так, в (12) топик, который первоначально, по-видимому, задумывался как топик-подлежащее, в ходе самоисправления был переквалифицирован в топик-косвенное дополнение:

(16) YAM R1:17–18

17. kaa san wa nani o ==
мама сан TOP что ACC

Мама что...

18. kaa san ni wa nani o kat-tara
мама сан DAT TOP что ACC купить-COND

Маме что купить?

Имеются и случаи сбоев, комплексно затрагивающих и морфологию, и синтаксис. В примере (13) первоначально задумывался переходный глагол, вероятно, *tomeru* ‘останавливать’. Оборванная форма *tom-m=* (с нефонологической долготой конечного согласного, свидетельствующей о хезитации) была забракована, а вместе с ней и прямое дополнение *me o* ‘глаза + Аккузатив’; в результате самоисправления говорящий выбирает непереходный глагол *tomaru* ‘останавливаться’ (словообразовательную пару к *tomeru* ‘останавливать’), а аккузативная форма заменяется на номинативную,т. е. вместо конструкции «остановить глаза (на)» строится конструкция «глаза остановились (на)»:

(17) TAK R1:6

6. ..(0.1) Soshite-e ma tokei ni me o
Тогда ну часы DAT глаз ACC

tom-m= || ..(0.4) tokei ni me ga
ост= часы DAT глаз NOM

tomar-imash-ita keredomo
остановить-ADR-PST хотя

Тогда, ну, [он свой] взгляд на часах ост= ... его взгляд остановился на часах…

Наконец, третьим типом операций при самоисправлении является отмена — говорящий заменяет забракованный фрагмент без сохранения его исходного значения или полностью отказывается от исходного замысла. Например, в (14) говорящий отменяет две неудачных попытки, прежде чем правильно квалифицирует расположение конечностей человека на картинке (правая/левая:рука/нога — возможно сбой происходит из-за того, что с человеком на изображении говорящий идентифицирует себя и не может быстро справиться с «зеркальным» пространственным дейксисом):

(18) ISI T2:29

29. ..(0.2) watashi wa ..(0.2) ee(0.7) …(0.7)
я TOP

migi || hi= || a migi ude to ..(0.2) migi
правый лев= а правый рука и правый

ashi || ee(0.2) hidari ude to
нога левый рука и

migi ashi ni gipusu o ham-e
правый нога DAT гипс ACC надевать.CNV

Мне на правую ... ле= правую руку и правую ногу... а! ...на левую руку и правую ногу наложили гипс и …

Типичными случаями отмены являются оговорки, при которых говорящий непроизвольно порождает незапланированный отрезок дискурса. Так, в (15) вместо запланированного «мой день рождения» говорящий произносит «день рождения мамы и папы» (по-видимому, имел место эффект прайминга — в предтексте о родителях уже шла речь):

(19) ISI T2:9–10

9. ....(2.3) chichi to haha no tanjoobi
папа и мама GEN день.рождения

dat-ta node ==
COP-PST потому.что

Так как был день рождения мамы и папы....

10. ..(0.2) a watashi no tanjoobi dat-ta
А! я GEN день.рождения COP-PST

node
потому.что

*А! [вернее,] так как был **мой** день рождения…*

Отмены встречаются в нашем корпусе реже двух других типов операций — они составляют 14,5 % от общего числа коррекций.

По четвертому критерию — линейному расположению забракованного фрагмента и его откорректированного коррелята — большинство (83 %) коррекций в нашем корпу е являются контактными, т. е. откорректированный коррелят следует непосредственно за забракованным фрагментом, см. де-

монстрационный пример (1). В более редких случаях (17 %) самоисправление происходит дистантно, как в (16), где говорящий преждевременно приступил к произнесению слова *jidoosha* ‘автомашина’, но понял, что «забыл» наречие, исправился и снова вернулся к произнесению этого слова:

(20) ISI R1:18

18. ..(0.4) sono ato watashi wa ji= || ee(0.4)
это после я TOP авт=

tyoodo ...(0.5) eeto ….(1.1) jidoosha no
как.раз это.самое автомашина GEN

saron no mae o torikakat-ta node
салон GEN перед ACC проходить-PST так.как

Так как я после этого мимо авт= как раз это самое... мимо автосалона проходил …

4. Сопоставление с данными по устной русской речи

Всего в нашем корпусе было зарегистрировано 88 эпизодов самоисправления, из которых он-лайн коррекций — 76 эпизодов, редактирования — 12 эпизодов. Разумеется, в силу небольшого объема корпуса приведенные данные могут дать лишь самое общее представление о количественном распределении отдельных типов самоисправлений, однако мы считаем полезным сравнить наши результаты с результатами, полученными по аналогичной методике для корпуса устных рассказов школьников о своих сновидениях (Подлесская, Кибрик 2006, 2009). Сравним прежде всего общую частоту коррекций: в нашем корпусе её среднее значение составляет 1,8 случая на 100 словоупотреблений, в русском корпусе — 2,9 случая на 100 словоупотреблений. В литературе имеются сходные данные по английским бытовым диалогам — около 3 случаев на 100 словоупотреблений (Shriberg 1994). Таким образом, порядок чисел сопоставим, что может свидетельствовать об универсальной природе явления. Относительно более низкая частотность самоисправлений в нашем корпусе может быть связана как с разницей в жанре дискурса (рассказы по картинкам vs. инициированные личные рассказы), так и с разницей в возрасте говорящих. Важным и пока малоизученным фактором является индивидуальная манера речи говорящего; данные по отдельным говорящим для корпуса «рассказов о сновидениях» в цитированных выше работах не указаны, по нашему японскому корпусу разброс по индивидуальным говорящим и по отдельным текстам у одного говорящего может весьма значительным — от полного отсутствия самоисправлений в отдельных рассказах до 13 эпизодов на 100 слов, при модальном значении 1,3 эпизода на 100 слов. Весь этот комплекс факторов нуждается в тщательной проверке.

Сравним теперь распределение по отдельным типам коррекций. Прежде всего, обращает на себя внимание расхождение в распределении макро- и микрокоррекций. В нашем японском корпусе, как было сказано выше, микрокоррекции составляют 93 %, макрокоррекции — 7 %. Данные по рассказам о сновидениях — 73,6 % микрокоррекций, 26,4 % макрокоррекций. Значительное численное превосходство микрокоррекций в обоих случаях очевидно, однако оно гораздо более выражено в русском корпусе. Можно предположить, что рассказ по картинке изначально предполагает достаточно жесткий каркас дискурсивной структуры, поэтому в них исправления, затрагивающие структуру дискурса, встречаются реже, чем в личном рассказе, однако нам кажется, что здесь задействован и другой — так сказать, «внутрилингвистический» фактор. Сравним распределение коррекций по объему забракованного фрагмента в «рассказах о сновидениях» и в японском корпусе. (Оценивая распределение по этому критерию, мы будем сравнивать наши данные с данными не по всему массиву самоисправлений в корпусе «рассказов о сновидениях», а только по массиву микрокоррекций, поскольку, как сказано выше, макрокоррекции в японском корпусе, в отличие от русского, крайне малочислены). Здесь наблюдается существенное расхождение: в «рассказах о сновидениях» среди микрокорреций исправления крупных фрагментов встречаются в два раза реже, чем исправления мелких, тогда как в японском корпусе они распределены практически поровну. Получается, что «недостаток» самоисправлений на межфразовом уровне (макрокоррекций) японский язык компенсирует преобладанием самоисправлений крупных фрагментов внутри клаузы (микрокоррекций), а русский язык наоборот — при самоисправлении чаще выходит за пределы клаузы, вместо того, чтобы исправлять крупные фрагменты внутри клаузы. Причины этого кроются, как нам кажется, в структурных различиях между русским и японским языком: раннее появление глагола в русской клаузе с самого начала проецирует дальнейшее её развертываение, поэтому самоисправление чаще приводит к слому всей конструкции, а в японском языке, где глагол располагается в абсолютном конце клаузы, говорящий может более «безнаказанно» корректировать достаточно протяженные куски клаузы, пока не артикулирован глагол. Естественно, это объяснение пока имеет статус гипотезы, нуждающейся в проверке.

Распределение по типу операции дало близкие результаты в японском и русском корпусах: повторы, модификации и обмены соотносятся в процентном отношении как 53,9 % : 31,6 % : 14,5 % в японском и как 60 % : 30,8 % : 8,3 % — в русском. Почти идентичным получилось и распределение по дистантности: кон-

тактное расположение забракованного компонента и откорректированного коррелята отмечается в 83 % в японском корпусе и в 84 % — в русском. По-видимому, распределение по этим двум критериям в большей степени определяется универсальными закономерностями речепорождения, а не лингвоспецифическими факторами. Однако, опять же, для статистически достоверных выводов нужна проверка на гораздо больших массивах текстов, распределенных по языкам, жанрам, возрасту и полу говорящего и проч.

В целом же, нам удалось показать системный характер лингвистических механизмов самоисправления, согласованность качественных и ряда количественных показателей в японской и русской монологической речи. Это позволяет говорить о том, что сведения о речевых сбоях имеют существенное значение для понимания универсальных языковых процессов, и следовательно их регистрация должна стать обязательным требованием к качественным корпусам устной речи.

Литература

1. *Подлесская В. И., Кибрик А. А.* Коррекция в устной русской монологической речи по данным корпусного исследования / Русский язык в научном освещении, 2006. №12(2). С. 7–55.
2. *Подлесская В. И., Кибрик А. А.* Речевые сбои и затруднения // Кибрик А. А., Подлесская В. И. (ред.) Рассказы о сновидениях: корпусное исследование устного русского дискурса. М.:ЯСК, 2009. С. 177–218
3. *Ohori Toshio.* Remarks on suspended clauses: a contribution to Japanese phraseology // Essays in Semantics and Pragmatics: In honor of Charles J. Fillmore. Eds. Masayoshi Shibatani and Sandra A. Thompson. John Benjamins, 1996. P. 201–218.
4. *Shriberg, E. E.* Preliminaries to a Theory of Speech Disfluencies. PhD dissertation. University of California at Berkeley, 1994.

Применение концептуальных сетей для выявления структуры семантической парадигмы прилагательных

Concept lattice implementation in semantic structuring of adjectives

Потемкин С. Б. (potemkin@philol.msu.ru)

МГУ им. М. В. Ломоносова, Россия

Рассматриваются методы формального анализа понятий (FCA) в применении к построению онтологических отношений в классе прилагательных русского языка, характеризующих внешность человека с привлечением WordNet. Выполнен анализ их семантической парадигмы на основе формального контекста, построенного с применением двуязычного словаря.

1. Введение

В настоящее время ведутся активные работы по созданию компьютерного тезауруса русского языка [1, 7, 8], аналогичного по своим структурным и функциональным возможностям распространенному тезаурусу WordNet [16]. Словари такого типа дают широкие возможности для отображения семантических отношений между значениями, получившими лексикализованное выражение в рамках некоторого языка. К сожалению, покрытие лексики подобными тезаурусами для языков, отличных от английского, остается ограниченным, несмотря на значительные усилия по расширению набора синсетов (основная семантическая единица WordNet; набор английских слов, которые в совокупности кодируют некоторое семантическое значение) и их взаимосвязей. Отсюда возникает необходимость автоматизированного получения лексико-семантических отношений из существующих источников, таких как тестовые корпуса или толковые словари. Для решения этой задачи привлекаются, в частности, методы формального анализа понятий (FCA) [11, 13, 14].

Нами разрабатываются методы использования двуязычных (англо-русских) словарей в качестве источника формального контекста и дальнейшего построения концептуальной сети для представления онтологических отношений в классе русских прилагательных.

2. Лексические источники

При определении структуры семантической парадигмы определенной группы слов необходимо опираться на возможно более полные лексические источники. Нами использованы:

- Общие и специальные англо-русские словари, лексическая база данных (ЛБД) [5].

В состав ЛБД включены англо-русские эквиваленты более чем из 30 общих и специальных словарей, включая 3-х томный словарь под редакцией Апресяна, словарь Мюллера, электронные словари Лингво, Полиглоссум, Промт и многие другие. Переводные словари подвергаются своего рода естественному отбору, поскольку они ежедневно используются переводчиками для практических целей, плохой словарь будет отвергнут и не выдержит переиздания;

- Оценка внешности человека (словарь) [2];
- WordNet [16];
- Толковые словари Ожегова, Евгеньевой, частотный словарь Шарова [9];

В настоящем докладе рассматривается семантическая парадигма на материале прилагательных, характеризующих внешность человека. Частотность рассматриваемой лексики весьма значительна: *большой* — 1631 ipm (число словоупотреблений на миллион), *хороший* — 854 ipm, *старый* — 528 ipm, *белый* — 493 ipm, [9] и т. д. Эта группа

выбрана также ввиду ее важности для уточнения системных отношений русской оценочной лексики, представлений о типах лексических значений, особенностей реализации коннотации, нормативных лексических ассоциациях[3], понимания структуры художественного произведения [6]. Важна она и для лингводидактики, как основа для создания различных пособий по развитию речи, обучения русскому языку как русских, так и иностранцев, а также для перевода художественной, юридической, психологической и др. литературы.

Представление значений имен прилагательных строится аналогично представлению других частей речи. Применяется компонентный анализ имен прилагательных с привлечением толковых словарей; корпусные исследования привлекаются для анализа сочетаемости в синтагмах типа прилагательное — существительное, что позволяет кластеризовать прилагательные как атрибуты, приписанные определенным существительным, для которых уже построена некоторая классификация[12]. Используются методы непосредственного полевого тестирования для выявления коннотаций, т. е. сужения круга возможных синтагматических партнеров (прилагательных) данной лексемы (существительного) [4]. Системные отношения в лексике описываются в тезаурусах, где прилагательному приписывается лексическое значение, часто в одном гнезде со сходным по семантике глаголом или существительным.

Представляется многообещающим для выявления семантики прилагательного использовать двуязычные словари и уже построенный иноязычный тезаурус типа тезауруса Роже или получившего в последнее время широкое признание WordNet. Отношения синонимии и антонимии в классе прилагательных достаточно хорошо разработаны, однако и в этой области привлечение двуязычных словарей существенно обогащает списки синонимов и особенно антонимов [5], ручной подбор которых требует значительных усилий. Другие типы отношений: гипонимия, меронимия, метонимия и пр. исследованы значительно меньше. Выявление указанных отношений между прилагательными представляет теоретический и практический интерес, особенно в применении к автоматической обработке текста. В данном случае непосредственная опора на структуру WordNet малопродуктивна. Достаточно сказать, что семантическая организация качественных прилагательных в WordNet полностью отличается от семантической организации существительных или глаголов. Прилагательные организованы в кластеры, каждый из которых привязан к «фокальному» прилагательному, имеющему антоним (к которому привязан свой кластер). Т.е. антонимия оказывается базовым семантическим отношением для кодировки значения прилагательных. Подобный подход объясняется тем фактом, что прилагательные выполняет атрибутивную функцию и что значительное число атрибутов являются биполярными. В классе прилагательных WordNet не построены иерархические отношения, подобные отношению гипонимии для существительных или тропонимии для глаголов и, как правило, не указывается прямой гипероним, вместо него дается ссылка «Pertains to noun …», то есть в классе прилагательных гиперонимом часто является имя существительное, например для прилагательных, обозначающих величину (*большой, малый, узкий, просторный*) родовым гиперонимом является существительное «размер». В настоящем исследовании, однако, мы будем стремиться отыскивать иерархические и др. связи, не выходя за рамки класса прилагательных.

3. Формальный анализ понятий (FCA, Formal Concept Analysis)

Формальный анализ понятий основан на интуитивном представлении о том, что понятие или концепт имеет две стороны: *экстент*, который содержит некоторые объекты, и *интент*, в который входят все атрибуты, свойственные этим объектам [16]. Для проведения формального анализа понятий необходимо, прежде всего, определить *формальный контекст* К:= (G,M,I), где G = множество объектов; M = множество атрибутов; и I = бинарное отношение между элементами G и M, показывающее, какие атрибуты m приписаны каждому из объектов g. Формальный контекст легко представить в виде таблицы. В Таблице 1 в качестве объектов фигурируют некоторые прилагательные русского языка, в качестве атрибутов — набор переводов этих прилагательных; определенное русское слово, напр. *алчный* имеет переводной эквивалент rapacious, на пересечении соответствующей строки и столбца поставлен крест.

Для определения *концепта* в формальном контексте вводится операция деривации → :

$$X \subseteq G: X \rightarrow X' : \{m \in M \mid gIm \text{ для всех } g \in X$$

$$Y \subseteq M: Y \rightarrow Y' : \{g \in G \mid gIm \text{ для всех } m \in X$$

В нашем примере пусть *X := {ХИЩНЫЙ, прожорливый}* и пусть *Y:={ravening, wolfish},* тогда *X' = {ravening, rapacious, ravenous}, Y' = {ХИЩНЫЙ, жадный}*, далее *X"= {ХИЩНЫЙ, жадный, прожорливый}*, и т. д. Можно показать, что в общем случае $X \subseteq X''$ и $X' = X'''$ а также $Y \subseteq Y''$ и $Y' = Y'''$

Формальным концептом в рамках данного формального контекста является пара (A, B), где A = B', B=A', т. е. A = множество объектов, каждый из которых имеет все атрибуты из множества B; B = множество атрибутов, каждый из которых приписан всем объектам множества A. Все формальные концепты для заданного формального контекста можно сформировать как (X", X') или (Y', Y"), перебирая все подмножества

$X \subseteq G$ или $Y \subseteq M$. Существуют алгоритмы для быстрого построения множества формальных концептов [15].

В нашей таблице выделены ячейки, представляющие формальный концепт (A, B); A= {*алчный, грабительский*}; B= {*rapacious, ravenous*} Формальные концепты данного контекста образуют частично упорядоченное множество $B(K)$, задаваемое отношением $\leq : (A_1, B_1) \leq . (A_2, B_2). \leftrightarrow A_1 \subseteq A_2 (\leftrightarrow B_2 \subseteq B_1)$. Это отношение называется отношением *субконцепта-суперконцепта* и $\leq$ определяет полную решетку $\underline{B}(K)$ на $B(K)$, которую можно изобразить в виде помеченного направленного графа (рис. 1). Вершинами графа являются формальные концепты, а ветви отражают отношение субконцепта-суперконцепта.

Для применения методов FCA к выявлению семантической парадигмы прилагательных русского языка предлагается использовать тщательно разработанный семантический тезаурус английского языка WordNet. Основной семантической единицей этого тезауруса является *синсет* — набор английских слов, которые в совокупности кодируют некоторое семантическое значение. Элементом синсета является WM — значение, выраженное отдельным словом (словосочетанием), входящим в синсет. Отдельное слово может входить в различные синсеты, что отражает полисемию, а также и омонимию, присущую данному слову. Между синсетами установлены отношения гипо-гиперонимии (для существительных), тропонимии (для глаголов), антонимии, меронимии и пр. Синсеты, содержащие прилагательные, как правило, не охвачены отношениями гипонимии, установление иерархических отношений между прилагательными затруднительно с теоретической и практической точек зрения [1,12]. Тем не менее, использование синсетов для выявления структуры семантической парадигмы прилагательных представляется возможным и многообещающим. Отметим, прежде всего, что двуязычный англо-русский словарь может эффективно применяться для расширения списка синонимов, а также определения семантической близости между синонимами русского языка [5]. Можно предположить, что взяв набор английских слов, входящих в синсет, $\{e_i\}$, т. е. синонимов с определенным значением, и выписав все их переводы на русский язык $L_j(e_i) = r_{ij}$, в пересечении $\cap_{ij} r_{ij}$ получим набор русских слов, кодирующих значение, эквивалентное значению синсета $\{e_i\}$. Вследствие различного членения действительности в английском и русском языке, которое является прямым отражением несовпадения способов категоризации и, следовательно, концептуализации атрибутивов, а также склонности англичан к большей детализации картины мира и номинации различных признаков, такое пересечение, как правило, оказывается пустым, либо содержит одно-два слова с очень широкой семантикой. Предлагается поэтому воспользоваться средствами FCA, которые позволят выявить всю структуру множеств $\{r_i\}_j$ в их взаимосвязи с синсетом $\{e_i\}$. Формальным контекстом K:= (G,M,I) в данном случае состоит из: множества объектов $G = \cup_j \{r_i\}_j$; всех переводов всех английских слов, входящих в синсет; множества атрибутов M = $\{e_i\}$; бинарного отношения I, определенного функцией L, сопоставляющей каждому английскому слову e_i его j-й русский эквивалент (Табл. 1).

4. Экспериментальные результаты и их интерпретация

В качестве массива данных для проведения экспериментальной апробации методики выбран Словарь «Оценка внешности человека» [2], (далее — Словарь), содержащий более 200 доминант и более 1200 членов синонимических рядов прилагательных, относящихся к внешности человека. В Словаре не выделены качественные и относительные прилагательные, грань между которыми во многих случаях весьма условна. В частности, к описанию лица относятся 603 прилагательных, для которых по изложенной ниже методике были построены 1040 концептуальных сетей с числом атрибутов более 2.

Для каждого прилагательного ar_i из Словаря отыскиваются все английские эквиваленты $ae_{ij}=L_j(ar_i)$, содержащиеся в лексической базе данных (ЛБД). Для каждого ae_{ij} определяется набор синсетов $\{s_k\} = WN(ae_{ij})$ содержащих ae_{ij}. Для каждого из синсетов s_k полученного набора выписываются все русские прилагательные, являющиеся переводными эквивалентами элементов данного синсета; дубли исключаются. Таким образом, получен набор объектов G и набор атрибутов M формального контекста K. На этом этапе мы не выполняем семантического разделения противоречивых переводных эквивалентов (которые в действительности встречаются, напр. *large-handed* переводится как *жадный* и как *расточительный*). Также не отбираются только прилагательные, относящиеся к внешности человека, такой отбор выполняется позже, на этапе анализа построенной концептуальной сети. Все возможные пары эквивалентов включаются в таблицу 1.

Таблица 1 Формальный контекст для синсета. Объекты, входящие в Словарь, выделены заглавными буквами

	edacious	esurient	ravening	rapacious	ravenous	voracious	wolfish
ЗВЕРИНЫЙ							×
ЗВЕРСКИЙ							×
СВИРЕПЫЙ							×
ХИЩНЫЙ			×	×	×		×
алчный				×	×		
грабительский				×	×		

	edacious	esurient	ravening	rapacious	ravenous	voracious	wolfish
волчий					×		×
голодный		×			×		
голодный_как_волк					×		
жадный	×	×	×	×	×	×	×
жаждущий					×		
захватнический				×			
изголодавшийся					×		
ненасытный		×		×	×	×	
относящийся_к_волкам							×
очень_голодный					×		
падкий						×	
похожий_на_волка							×
прожорливый	×	×	×	×	×	×	
свинский				×			
характерный_для_волка							×
эгоистичный				×			

Рис. 1. КС для синсета №00011320 {edacious, esurient, rapacious, ravening, ravenous, voracious, wolfish} к которой относятся объекты {ЗВЕРИНЫЙ, ЗВЕРСКИЙ, СВИРЕПЫЙ, ХИЩНЫЙ}

На рисунке 1 показана концептуальная сеть для формального контекста таблицы 1. В рамках синсета №00011320 объект ХИЩНЫЙ выявляется как гипероним для объектов ЗВЕРИНЫЙ, ЗВЕРСКИЙ, СВИРЕПЫЙ. Такое определение гиперонима в общем случае не представляется корректным (зверь не обязательно хищник, см. Ефремова: *зверь1 = Дикое, обычно хищное животное*), но в качестве характеристики лица звериное, зверское, свирепое лицо скорее всего есть лицо хищное. Из рассмотрения других синсетов выявлены следующие отношения гипонимии:

мертвый ⊆ неподвижный ⊆ вялый
апатичный, оцепенелый ⊆ вялый
изящный ⊆ тонкий
коварный ⊆ хитрый
нахальный, самоуверенный ⊆ дерзкий ⊆ смелый
решительный ⊆ твердый
ястребиный ⊆ хищный
мерзкий, отвратительный, противный, ужасный ⊆ неприятный

Некоторые из этих отношений совпадают с приведенными в Словаре (*изящный ⊆ тонкий, коварный ⊆ хитрый),* остальные выявлены вновь, либо противоречат Словарю, напр. в качестве гиперонима к *ястребиный* в Словаре указано прилагательное *беличий* (?).

Кроме выявления отношений гипонимии из концептуальных сетей можно извлечь прилагательные, которые могли бы войти в Словарь: *бесчувственное, будничное, выцветшее, загадочное, заспанное, зловещее, искаженное, легкомысленное, матовое, незамысловатое, нездоровое, неприметное, плоское, подозрительное, полное, полусонное, придурковатое, притворное, разбойничье, смущенное, сухощавое, флегматичное, худое, худощавое,…*

Также выявлены словосочетания, выполняющие атрибутивную роль, которые вообще не включены в словники Словаря: *с буйной растительностью, наводящее скуку, с хитрецой, с хитринкой…* Сопоставление всех полученных иерархических отношений со Словарем не является задачей данного исследования. Предложенный метод позволил выявить дополнительные лексические единицы и установить семантические связи, которые могут использоваться как в лексикографии, так и для решения задач АОТ.

5. Выводы и перспективы исследования

Сложность задачи выявления семантической структуры класса прилагательных подтверждена проведенными ранее исследованиями. Применение методов формального анализа понятий (FCA) для ее решения может оказаться полезным в качестве дополнения к методам корпусных исследований, компонентного анализа и др. Предполагается развить описанные методы с целью формализации выделения иерархических отношений в построенной концептуальной сети. Кроме того, возможно распространение приведенного подхода на иные семантические отношения.

Литература

1. *Азарова И. В., Синопальникова А. А., Яворская М. В.* Принципы построения wordnet-тезауруса RussNet // Компьютерная лингвистика и интеллектуальные технологии. Труды Международной конференции Диалог'2004 М.: 2004. С. 542–547.
2. *Богуславский В. М.* Оценка внешности человека, словарь изд-во «Аст» М. 2004г. 255 стр.
3. *Кедрова Г. Е., Потемкин С. Б.* Семантическое разделение омонимов с использованием двуязычного словаря и словаря синонимов // Труды II Международного конгресса «Русский язык: исторические судьбы и современность», М. 2004 г.
4. *Кобозева И. М.* Лингвистическая семантика изд-во «Эдиториал УРСС», М. 2000г. 350 стр.
5. *Потемкин С. Б.* Лексическая база данных с наложенной семантической метрикой // Труды II Международного конгресса «Русский язык: исторические судьбы и современность», М. 2004 г.
6. *Потемкин С. Б.* Обнаружение события путем анализа антонимов в текстах Н. В. Гоголя и А. П. Чехова, // Слово и словарь — Труды Международной научной конференции «Современные проблемы лексикографии», Гродно 2009 С. 93–95
7. *Портал УИС «Россия»* http://www.cir.ru/.
8. *Сухоногов А. М. Яблонский С. А.* Автоматизация построения англо-русского WordNet. //Труды RDCL 2004. 29 сентября — 1 октября. Пущино, 2004 г.
9. *Шаров С. А.* Частотный словарь http://www.artint.ru/projects/frqlist.asp
10. *Яворская М. В., Азарова И. В.* Структура атрибутивных значений в тезаурусе RussNet (на материале перцептивных прилагательных) // Труды Международной конференции Диалог'2009 С. 542–547.
11. *Cimiano P., Hotho A., Staab S.* Learning Concept Hierarchies from Text Corpora using Formal Concept Analysis. // Journal of Artificial Intelligence Research. Volume 24, August 2005 P. 305–339
12. *Mendes S.* Adjectives in WordNet.PT // GWC 2006, Proceedings, P. 225–230.
13. *Priss U.* Linguistic Applications of Formal Concept Analysis. // Ganter; Stumme; Wille (eds.), Formal Concept Analysis, Foundations and Applications. Springer Verlag. LNAI 3626, 2005, P. 149–160.
14. *Stepanova N. A.* Automatic acquisition of lexico-semantic knowledge from corpora. // SENSE'09 Workshop pp. P. 91–100, 2009
15. *Wille R.* Restructuring lattice theory: an approach based on hierarchies of concepts. // Rival, I. (ed.) Ordered Sets. 445–470. Dordrecht-Boston, Reidel, 1982.
16. *WordNet:* An Electronic Lexical Database // Fellbaum Ch. (ed.). MIT Press. 1998.

Система подготовки нового голоса для системы синтеза «VitalVoice»

Voice building system for hybrid Russian TTS system «VitalVoice»

Продан А. И. (prodan@speechpro.com),
Таланов А. О. (andre@speechpro.com),
Чистиков П. Г. (chistikov@speechpro.com)

ООО «Центр речевых технологий», (Санкт-Петербург, Россия)

Рассматривается технология создания нового голоса заданного диктора для работы в системе синтеза VitalVoice. Описана система автоматизированной подготовки голоса, выбор текстового материала, особенности процесса записи речи, создание базы Unit Selection, настройка параметров подбора элементов.

1. Введение

Система автоматизированной подготовки голоса используется при подготовке речевых данных для синтезатора речи по тексту VitalVoice ООО «Центр речевых технологий» (ЦРТ) [4, 11, 12].

Существуют различные подходы к организации автоматического синтеза речи по тексту. К основным можно отнести синтез по правилам (формантный синтез), артикуляторный синтез, компилятивный синтез, синтез на основании статистических моделей (HMM-синтез). Синтез методом Unit Selection [2, 3] — один из видов компилятивного синтеза. Его отличительной особенностью является то, что синтезированная речь составляется не из базы специально записанных аллофонов, дифонов или других элементов, каждый из которых представлен единственным вариантом, а из элементов, взятых из естественных предложений, и для каждого элемента производится выбор наиболее подходящего кандидата из множества вариантов. Данная технология позволяет достичь очень высокой естественности синтезированной речи. В рамках работы по созданию новой системы синтеза русской речи, осуществляемой ЦРТ, создан синтезатор на основе использования технологии Unit Selection, совмещенной с аллофонным синтезом.

Характерной особенностью синтеза методом Unit Selection является его критическая зависимость от состава и полноты речевого корпуса. Качественный синтез определённым голосом возможен только на основе полного, сбалансированного и корректно размеченного речевого корпуса. В ЦРТ для разметки речевой базы для синтеза Unit Selection была разработана специальная многоуровневая система [13].

Задача добавления нового голоса, безусловно, является очень актуальной для любой системы синтеза речи. В особенности это актуально для синтеза методом Unit Selection, поскольку для этого метода это крайне ресурсоемкая задача. Именно поэтому она является предметом разных направлений исследований. С одной стороны, она интересна с точки зрения изучения тех или иных характеристик голоса и речи определённого диктора, которые влияют на качество синтезированной речи, с другой — с точки зрения задачи максимального приближения синтезированной речи по своим характеристикам к оригинальной речи самого диктора, то есть имитации. Отдельной интересной подзадачей является создание голоса по уже имеющемуся речевому материалу без участия диктора в записи (например, имитация голоса известных актёров).

Поэтому очень важно сделать процесс подготовки голоса по возможности максимально быстрым и удобным. Причём инструмент, помогающий в подготовке звуковой базы и настройке голоса, должен подходить как для специалиста (то есть подразумевает наличие ручных настроек всех процессов, их корректировки и т. п.), так и для человека, имеющего только самые общие представления о фонетике и синтезе речи (или получившего их из прикладываемой к программе документации и пошаговой инструкции), — то есть свести ручную корректировку и настройку к минимуму, в идеале к «нажатию одной кнопки».

Система подготовки нового голоса (СПГ) предназначена для автоматизации работы по созданию голоса для системы синтеза VitalVoice. В результате работы СПГ формируется установочный файл голоса, который работает с программой синтеза русской речи VitalVoice.

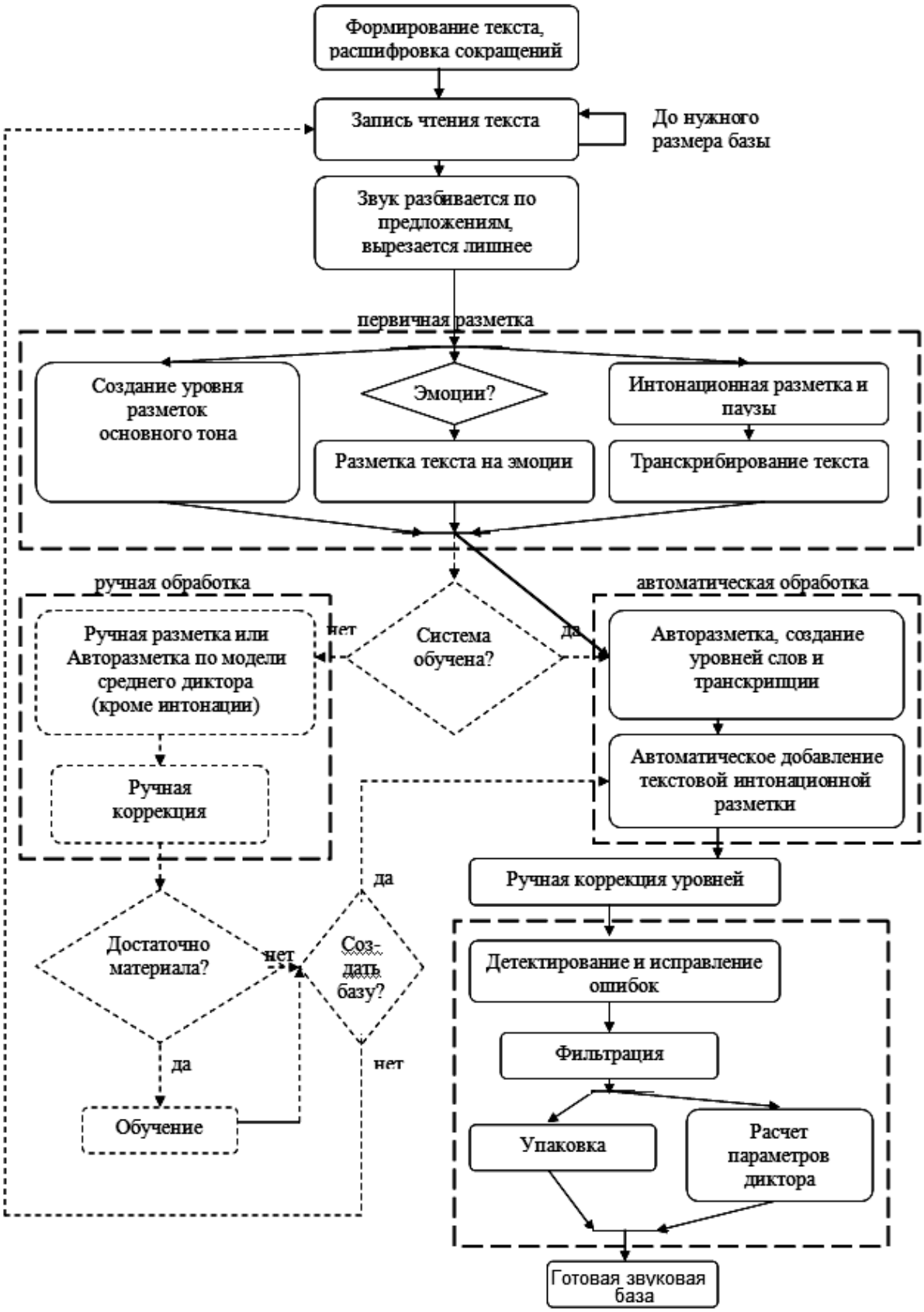


Рис. 1. Алгоритм создания звуковой базы

Система состоит из следующих частей:

- сама система подготовки голоса;
- звуковой редактор;
- транскриптор;
- программа автоматической разметки;
- программа автоматизированной проверки разметки базы;
- упаковка базы, создание установочного файла;
- расчёт и настройка параметров диктора;
- пошаговая инструкция по созданию и настройке голоса.

2. Основные шаги СПГ

Процесс создания звуковой базы представляет собой последовательность взаимосвязанных и при необходимости повторяющихся действий, объединённых в блоки, которые в разной степени поддаются автоматизации. На рис. 1 изображён алгоритм создания звуковой базы Unit Selection для синтеза VitalVoice.

Автоматизированное создание голоса происходит посредством выполнения последовательности шагов. На каждом шаге есть возможность перейти к следующему с использованием параметров по умолчанию (если не возникло ошибок, в противном случае — исправить только критические или все). Для исполнения каждого шага программа предложит совершить определённые действия. Стандартное окно программы изображено на рис. 2.

На каждом шаге предлагается его краткое описание и ссылка на подробную инструкцию. Рабочий экран содержит элементы управления, необходимые для выполнения действий данного этапа. При открытии каждого этапа кнопка «Дальше» остается недоступной до нажатия кнопки «Принять» и успешного прохождения всех проверок.

Флаг «автозапуск процессов» служит для установки автоматического запуска процессов обработки, предусмотренных на этапе, запуска проверки полученных результатов и перехода к следующему этапу в случае успеха.

2.1. Запись звуковых файлов

На этом этапе записывается чтение диктором выбранного текста. Если звуковые файлы уже записаны или необходимо создать голос по уже имеющимся записям, указывается текст, который был предложен диктору (если он есть, иначе текстовую расшифровку можно сделать на следующем этапе), и соответствующие ему звуковые файлы.

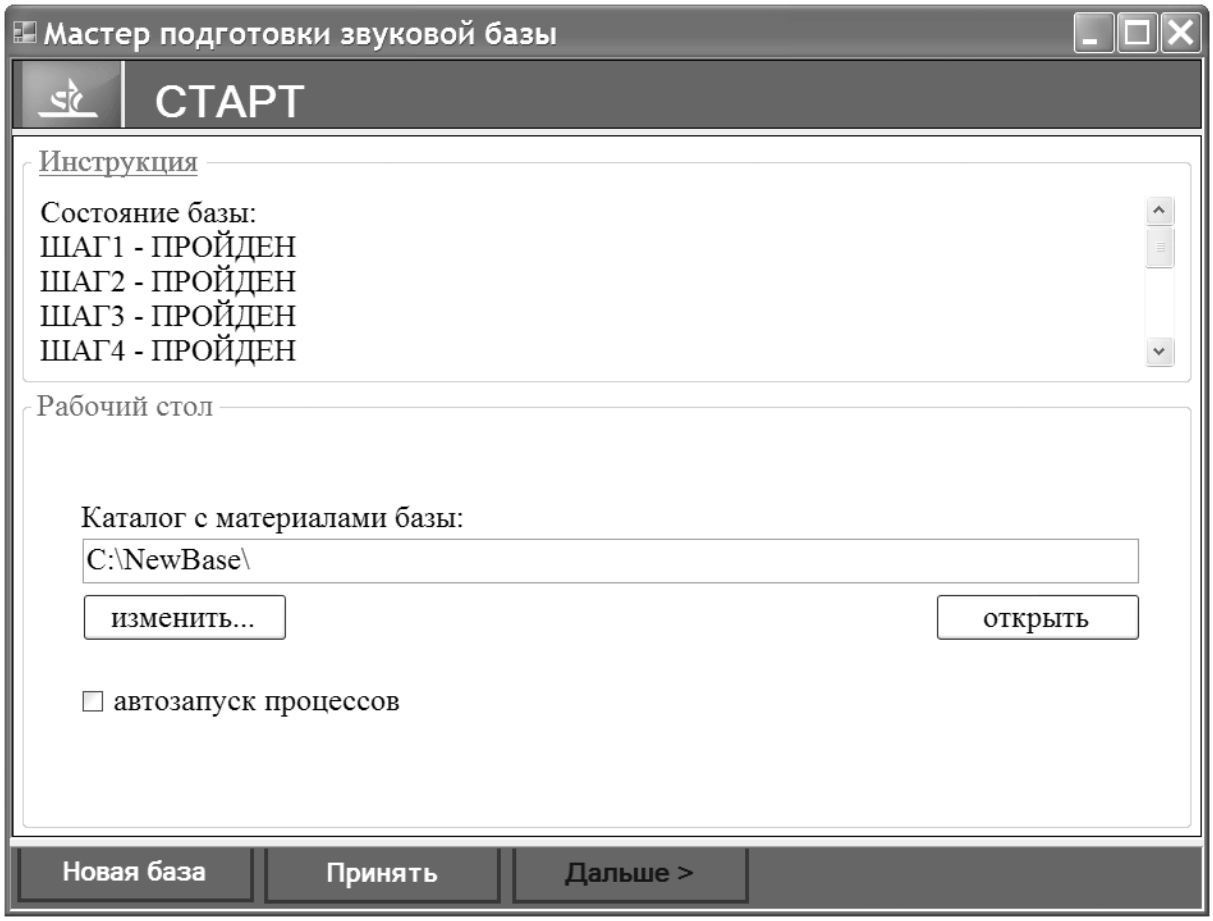


Рис. 2. Основная форма приложения, шаг «Старт»

2.1.1. Требования к тексту

Для создания качественного синтезированного голоса методом Unit Selection важно, чтобы на минимальном объёме звучащей речи было представлено максимальное количество элементов, по которым ведётся поиск, в системе синтеза VitalVoice это аллофоны в разных контекстах (т. е. трифоны) [1, 5]. Если минимальной единицей является не аллофон, а дифон, часть аллофона и т. д., важно, чтобы база была представительной с этой точки зрения. То же требование касается максимально возможной представительности интонационных конструкций. В «Центре речевых технологий» были разработаны специальные инструменты, позволяющие получить статистику наличия трифонов в тексте, а также подобрать фразы, содержащие аллофоны в редком контексте. Если в базе есть интонационная разметка, то по ней также можно получить статистику по использованию типов ИК [9, 10] и их параметрам. Кроме аллофонной и интонационной вариативности в текст также полезно добавить частотные слова, числа, фразы и, особенно если синтез будет использоваться для какой-то определённой задачи, названия и фразы, характерные для этой сферы (например, общение с системой голосового самообслуживания).

2.1.2. Условия записи

Запись диктора желательно производить в студии с хорошей шумоизоляцией — уровень сигнал/шум должен быть не меньше 30–40 дБ. Рекомендуется использовать конденсаторный микрофон. Запись наших баз производилась с частотой дискретизации 22 050 Гц и разрядностью 24 бита. При записи необходимо придерживаться уровня громкости −12 … −10 дБ.

Чтение диктором текста

Текст лучше читать в нейтральной, сдержанной манере. Реализации различных интонационных конструкций не должны отличаться сильным разнообразием. Манера речи должна быть спокойной, доброжелательной. Темп и ритм чтения ровные. Между предложениями обязательны паузы. Лучше делать их больше естественных, чтобы можно было разбить запись по предложениям автоматически без ручной корректировки, настроив нужную длину паузы для разбиения. Внутри предложений частые паузы также излишни.

Желательно следить за чёткостью артикуляции всех звуков, но не делать её нарочитой и неестественной. Как правило, типичные для диктора темп и тембр устанавливаются после нескольких первых минут записи. При начале каждой новой сессии, особенно если прошлая запись происходила в другой день, надо сверять тембр и темп с предыдущими записями, полезно также дать послушать их диктору. В работе обязательны небольшие перерывы, а если диктор устал, запинается или у него изменился тембр голоса — лучше продолжить запись в другой день. Ни в коем случае нельзя записывать диктора с простудой или охрипшим голосом. При записи желательно постоянно работать с диктором, т. е. непрерывно следить за темпом и тембром и поправлять его.

При выборе готовых записей для создания голоса следует отдать предпочтение чтению перед спонтанной речью (интервью и т. п.), а также обратить внимание на зашумлённость записи и реверберацию. Те участки, где на фоне речи целевого диктора присутствует речь других людей, полностью удаляются.

2.2. Разбивка звуковых файлов по предложениям

На этом шаге звуковые файлы, содержащие озвученный текст, разбиваются на части, содержащие по одному предложению или синтагме, в соответствии с текстом (или просто по паузам). Некачественный материал (зашумлённые участки, оговорки, речь других людей и повторы, если они по каким-либо причинам не нужны) удаляется. Удаляются длинные паузы в начале итоговых звуковых файлов. Также на этом этапе удобно делать текстовую расшифровку записи, если диктор не читал заранее подготовленный текст.

Разбить исходный звуковой файл можно автоматически. Если разбивка прошла неудачно (слишком короткие/длинные куски), изменяются в соответствующую сторону параметры паузы, максимальная амплитуда и минимальная длина.

Необходимо синхронизировать предложения в тексте и пронумерованные звуковые файлы. Лишние файлы удаляются. Слишком короткие можно объединить.

Для удобства все предложения в тексте нумеруются. При переходе на следующий шаг автоматически проверяется наличие соответствующих номерам предложений звуковых файлов и наоборот, а также, исходя из среднего темпа чтения, при сильных отклонениях выдаются предупреждения о слишком коротких или длинных звуковых файлах, если текст намного короче или длиннее звука соответственно. На этом подготовительном этапе участие оператора для проверки и синхронизации звуковых файлов и текста необходимо, так как в ином случае будут неизбежны ошибки в разметке и составлении базы.

2.3. Разметка на периоды основного тона

На этом шаге для каждого файла, отобранного на предыдущем этапе, делается разметка на периоды основного тона. Это можно сделать с помощью программы WaveAssistant вручную или автоматически в режиме пакетной обработки.

Для более высокого качества синтеза и автоматической разметки на аллофоны желательно после автоматической разметки на периоды ОТ производить ручную правку тона для каждого файла.

2.4. Транскрибирование текста

Для каждого предложения автоматически создаётся файл с транскрипцией и текстовой расшифровкой в специальном формате для программы автоматической разметки. Правила транскрипции и правила лингвистической обработки большей частью задаются во внешних текстовых файлах, при необходимости (например, чтобы учесть какие-либо индивидуальные особенности диктора) в них можно оперативно внести изменения. Изначально вся обработка до стадии получения транскрипции совпадает с обработкой текста программой синтеза речи [8, 14], но вынесена в отдельное приложение, в свою очередь встроенное в систему подготовки нового голоса.

2.5. Авторазметка на аллофоны

Разметка на аллофоны — основной шаг при подготовке нового голоса.

Разметка может быть произведена как вручную, так и при помощи программы автоматической сегментации, разработанной в «Центре речевых технологий» на основе компонентов системы автоматического распознавания речи. При помощи неё создаются уровни идеальной, реальной транскрипции и уровень слов.

На уровне идеальной транскрипции расставляются метки аллофонов, полученные на шаге транскрибирования текста. На уровне реальной транскрипции выполняется пофонемное распознавание, при окончательном выборе аллофона учитывается его соответствие аллофону на уровне идеальной транскрипции. Возможные варианты аллофонов идеальной транскрипции или их пропуск задаются в подгружаемой из внешнего файла таблице соответствий.

Более точной является разметка, основанная на акустических моделях, построенных на основе ручной сегментации речи заданного диктора на аллофоны, но для этого требуется размеченная база голоса объёмом не менее полутора часов. Если такого объема ручной сегментации нет, то программа автоматической разметки будет использовать акустические модели, обученные по любым другим речевым базам (совпадающим по техническим характеристикам сигнала). При этом в процессе обработки других баз будут использованы только дикторы с похожими голосами. Тот же принцип программа авторазметки использует в случае нехватки звукового материала при построении акустических моделей редких трифонов.

2.6. Коррекция ошибок разметки

На этом этапе производится проверка полученной на прошлом шаге автоматической разметки. Для синтеза речи заданным голосом с высоким качеством звучания желательно выполнять проверку разметки каждого файла как вручную, так и автоматически. В систему автоматической проверки разметки звукового корпуса [13] были введены некоторые новые проверки, относящиеся в основном к анализу автоматической разметки: проверка аллофонов, выбивающихся по длительности, более жёсткая проверка соответствий реальной и идеальной транскрипции и т. д. Кроме этого, ряд ошибок, препятствующих сборке базы голоса, выделен в отдельный блок, причём переход на следующий шаг осуществляется только после их исправления. Остальные блоки ошибок отсортированы по степени важности, и решение об их исправлении выносится в зависимости от наличия времени и ресурсов.

2.7. Фильтрация звука

На этом шаге каждый звуковой файл базы фильтруется. Это делается для того, чтобы по возможности максимально избежать при стыковке эффекта реверберации, то есть в основном следов предшествующих гласных на глухих согласных и звонких смычных. Глухие согласные фильтруются ФВЧ с полосой задерживания от 1500 Гц, звонкие после гласных с большой амплитудой — ФНЧ с полосой задерживания от 450 Гц.

2.8. Упаковка базы, расчёт параметров диктора и создание установочного файла голоса

Все необходимые для упаковки файлы к этому этапу уже готовы. Упаковка базы производится автоматически. Во время сборки подсчитываются различные характеристики (средняя длительность, амплитуда аллофонов, средний основной тон диктора и т. д.), которые потом можно сразу использовать как параметры подбора аллофонов для диктора. Если была произведена интонационная разметка, можно также заменить стандартные настройки типичными для диктора.

Настройка параметров подбора аллофонов

Параметры подбора элементов [6, 7] для лучшего качества синтеза зависят как от характеристик речи диктора и точности разметки базы, так и от размера базы. По умолчанию выбираются параметры для средней базы диктора заданного пола.

Далее приводятся общие рекомендации по параметрам подбора элементов. Чем меньше размер

базы, тем больший вес следует поставить на соединение по тону и спектру, но меньше вес на неразрывность последовательности аллофонов. В базе небольшого объёма больше вероятность того, что аллофон в целевом контексте найден не будет, поэтому нужна большая свобода выбора неточного контекста. Но, с другой стороны, если разметка неточна, то есть границы аллофонов сильно смещены, этот вес снижать не рекомендуется. В целом, если более приоритетной является естественность голоса, следует увеличивать стоимость связи, а если важнее точная передача интонации и максимальная разборчивость в ущерб естественности — стоимость замены.

3. Выводы и перспективы

На данный момент в системе синтеза VitalVoice десять основных голосов (четыре мужских и шесть женских), созданных на базах разного размера — от полутора до восьми часов звучащей речи. Опыт создания голосов подтверждает, что сбалансированной базы размером в полтора-два часа достаточно для довольно качественного синтеза, но для максимальной естественности желательно увеличить объём до 6–7 часов в зависимости от характеристик диктора.

Автоматизированная система создания голоса на текущий момент была опробована на небольших объёмах речевого материала (менее получаса), но уже позволила получить практически важные результаты: с минимальной ручной корректировкой разметки достигнута почти полная разборчивость речи и практически стопроцентная узнаваемость исходного диктора.

Развитие системы предусматривает как развитие каждого отдельного компонента, так и улучшение их взаимодействия и интеграции в системе подготовки голоса. В будущем планируется уделить наибольшее внимание автоматической разметке, дальнейшей автоматизации настройки параметров подбора элементов в зависимости от характеристик речи диктора и размеров базы, а также удобству использования самой программы.

Литература

1. *Black A. W.* Perfect Synthesis for all of the people all of the time // Keynote, IEEE TTS Workshop Santa Monica, CA, 2002.
2. *Black A. W., Hunt A. J.* Unit Selection in a Concatenative Speech Synthesis Using a Large Speech Database // In Proceedings of ICASSP 96. Atlanta, Georgia, 1996. Vol. 1, pp. 373–376.
3. *Clark R. A. G., Richmond K., King S.* Multisyn: Open-domain unit selection for the Festival speech synthesis system // Speech Communication, 2007. Vol. 49, issue 4. P. 317–330.
4. *Oparin I., Talanov A.* Outline of a New Hybrid Russian TTS System // Proc. of the 12th International conference on Speech and Computer, SPECOM 2007, Moscow, Russia, 2007. P. 603–608.
5. *Tatham M, Morton K.* Developments in Speech Synthesis // John Wiley & Sons Ltd, 2005.
6. *Vepa J., King S.* Subjective evaluation of join cost functions used in unit selection speech synthesis // In Proceedings of the International Conference on Speech and Language Processing 2004. Jeju, Korea, 2004. P. 1181–1184.
7. *Vepa J.* Join Cost for Unit Selection Speech Synthesis // University of Edinburgh, 2004.
8. *Аничкин И. М., Чистиков П. Г.* Формализация правил автоматического снятия омонимии в системе синтеза речи по тексту // Материалы XXXVIII Международной филологической конференции.
9. *Брызгунова Е. А.* Интонация // Русская грамматика. М.: 1980.
10. *Вольская Н. Б., Скрелин П. А.* Моделирование интонации для синтеза речи по тексту // Уфа: 1998.
11. *Киселёв В. В., Чижденко В. А., Таланов А. О., Опарин И. В.* Архитектура системы синтеза русской речи по тексту нового поколения // Материалы XXXVII Международной филологической конференции.
12. *Корольков Е. А., Главатских И. А., Таланов А. О., Киселев В. В., Опарин И. В.* Синтез естественной русской речи при помощи метода Unit Selection // Материалы XXXVII Международной филологической конференции.
13. *Продан А. И., Корольков Е. А., Опарин И. В., Таланов А. О.* Особенности использования многоуровневой разметки звукового корпуса Unit Selection в системе гибридного синтеза «Живой голос» // Материалы международной конференции Диалог 2009.
14. *Хомицевич О. Г., Рыбин С. В., Таланов А. О., Опарин И. В.* Автоматическое определение места ударения в незнакомых словах в системе синтеза речи // Материалы XXXVII Международной филологической конференции.

Теория и реальность: номинализация глаголов в разговорной речи[1]

Theory and reality: nominalization of verbs in spoken language

Розина Р. И. (rarozina@yandex.ru)

Институт русского языка им. В. В. Виноградова РАН, Москва

Рассматривается проблема наследования отглагольными именами семантики и стилистической окраски разговорных и сленговых глаголов. Вводится понятие степени разговорности глагола. Исследуется полисемия отглагольных имен и различия деривации литературных и сленговых производных значений.

Существует распространенное мнение о том, что отглагольные имена имеют книжную стилистическую окраску, и их использование — характеристика официально-делового стиля. Задача данной статьи — проверить, так ли это на самом деле, рассмотрев отглагольные имена, образованные от разговорных и сленговых глаголов русского языка.

Различным аспектам номинализации, в первую очередь проблеме наследования отглагольным именем модели управления мотивирующего его глагола, посвящена огромная лингвистическая литература (см., например, библиографию в работах [Пазельская 2003], [Пазельская 2005], [Цзя Хуа Чжан 2007], [Rozwadowska 1997] и др.), но номинализация разговорных и сленговых глаголов никогда не попадала в поле зрения исследователей. Единственное исключение — статья [Розина 2008], в которой анализируются модели деривации разговорных отглагольных имен, связь между их семантикой и семантикой мотивирующего глагола и актантные характеристики отглагольных существительных, в частности возможность выражения второго, объектного актанта. В статье доказывается, что по ряду причин актантная структура мотивирующего глагола наследуется разговорными или сленговыми отглагольными именами значительно реже, чем образованными от этих же или других глаголов отглагольными именами в литературном языке.

В центре внимания в данной работе — проблема наследования отглагольным существительным стилистических и семантических характеристик мотивирующего глагола. Мы покажем, что стилистическая окраска отглагольного имени, мотивированного разговорным или сленговым глаголом, зависит от ряда взаимодействующих факторов и что семантическая структура отглагольного имени не является точным отображением семантической структуры мотивирующего глагола.

В основе исследования, результаты которого излагаются в статье, лежат материалы Национального корпуса. В отдельных случаях, когда примеров использования слова в Корпусе было очень мало или они отсутствовали вообще, использовались данные блогов, полученные с помощью поисковой системы «Яндекс».

Источником выборки разговорных и сленговых глаголов послужил словник Толкового словаря разговорной лексики[2] . В словник вошли слова, имеющие помету *разг.* в МАСе, в словаре [Шведова 2007] и ряде других толковых словарей, а кроме того слова из ряда словарей разговорной речи, словарей новых слов и значений и словарей сленга и жаргонов, например [Ермакова, Земская, Розина 1999], [Химик 2004], [Курилова 2006]. Из словника подряд по алфавиту отбирались глаголы, способные к номинализации, независимо от того, включались в словник соответствующие

[1] Работа поддерживается грантами РГНФ № 08-04-00181а.

[2] Словарь составляется коллективом сотрудников Отдела современного русского языка Института русского языка им. В. В. Виноградова РАН под руководством Л. П. Крысина. Словник составлен Л. П. Крысиным и обсужден и дополнен этим же коллективом.

им отглагольные имена или нет[3]. Всего было отобрано двадцать шесть глаголов и образованных от них отглагольных имен: *аукать — ауканье, ахать — аханье, бахвалиться — бахвальство, баюкать (ребёнка) — баюканье, болтать1 (самолет)- болтанка, болтать2 (о чем-л.) — болтовня, брехать — брехня, вилять — виляние, ворковать — воркотня, впарить (зонтик кому-л.) — впаривание, долбить (уроки) — долбёжка, заварить (кашу) — заваруха, заводить (всех вокруг себя) — заводка, заделать (щели) — заделка, зажимать (свободу прессы) — зажим, заказать (кого-л. убить) — заказ, заказуха, закатывать (помидоры) — закатка, закачать (файл) — закачка, заморачивать (голову) — заморочка, замочить (террористов) — замочка, засветить (Киркорова) — засветка, засыпать (студента на экзамене) — засыпка, заточить (кого-л./ что-л. на/под что-л.) — заточка, зачистить (деревню от боевиков) — зачистка, крышевать — крышевание, отмывать (деньги) — отмывание*.

1. Что такое разговорный глагол?

Множество глаголов, которые в словарях объединяются пометой «разг.», крайне разнородно. Вопрос о включении того или иного глагола в словник словаря разговорной лексики часто вызывает ожесточенные споры. Поэтому, прежде, чем рассматривать стилистическую окраску имен, производных от разговорных и сленговых глаголов, стоит попытаться понять, что может скрываться за такой характеристикой, как разговорность.

Общеизвестно, что четких критериев «разговорности» не существует. Кажется разумным на примере нескольких типично разговорных лексических единиц, в данном случае глаголов, выявить ряд их характерных особенностей, которые затем можно будет применять в качестве диагностических критериев при решении вопроса о стилистической окраске остальных глаголов.

Пример типичного разговорного глагола, т. е. такого глагола, который признается разговорным всеми носителями языка, — глагол *врубиться* ‘понять’. В число особенностей этого глагола входят:

а) наличие нейтрального синонима — *понять*;

б) метафорический характер значения (перенос по модели ‘физическое действие — ментальное действие’)

в) возможность использования в разговорном контексте, в частности в диалогической речи — ср.:

(21) *Сунул в кармашек и сказал, пьяно усмехаясь: — Серега, а зачем ты мне штуку отдал, я что-то не* ***врублюсь****…* [Андрей Волос. Недвижимость (2000) // Новый Мир, № 1–2, 2001].

г) неуместность в официальном контексте, ср.:

(22) *Я не ****врубился*** *в то, какими правовыми принципами руководствовались члены совета FIS, принимая решения о дисквалификации*[4].

Глагол *сжевать* в значении ‘съесть’ обладает теми же особенностями, что глагол *врубиться* (имеет нейтральный синоним съесть, используется исключительно в разговорных контекстах), но отличается от него характером значения: у *сжевать* оно не метафорическое, а метонимическое, ср.:

(23) *Смогла только влить в себя утренний кофе (без него я не жилец) и* ***сжевать*** *два финика* (блоги).

Иными словами, общий признак глаголов *врубиться* и *сжевать* — переносный характер значения.

Такие разговорные глаголы, как *пахать, батрачить* ‘тяжело работать’, *долбить* ‘учить’, *влететь* и *попасть* ‘оказаться в неприятном положении’, *пролететь* ‘потерпеть неудачу’ позволяют добавить к перечисленным признакам еще один: наличие оценочного (в случае этих глаголов — отрицательного) компонента в значении лексемы.

Рассмотрим теперь глагол *зазнаться*, который и в МАСе, и в Большом толковом словаре синонимов русской речи (Бабенко 2008) имеет помету *разг.* У этого глагола нет нейтрального синонимичного слова: в МАСе он толкуется через устаревший разговорный глагол *возомнить (о себе)* и через словосочетание *стать высокомерным*; В словаре синонимов под ред. Бабенко все глаголы, синонимичные глаголу *зазнаться / зазнаваться*, кроме *превозноситься / превознестись*, снабженного пометой устар. или ирон., имеют помету разг. У глагола *зазнаться / зазнаваться* есть только одно значение — тем самым оно не может быть метафорическим. Иными словами, из пяти перечисленных признаков у этого глагола есть только три: в его значении есть отрицательный компонент (‘зазнаваться — это плохо’), он употребляется в разговорном контексте и не употребляется в официальной речи.

---

[3] В некоторых случаях отглагольное имя, соответствующее разговорному или сленговому глаголу, в словнике отсутствует — например, потому что оно было исключено из-за своей книжной стилистической окраски. Поскольку нам интересен любой результат номинализации, мы вводили это имя в свой список. Так, например, в список было добавлено отглагольное имя *бахвальство* в пару к глаголу *бахвалиться*.

[4] В оригинальном тексте …*не понял то*… [Андрей Митьков. По пути в Гаагу. «Дело Лазутиной» дошло до гражданского суда. За это спортивные чиновники грозят дисквалифицировать всю Россию (2002) // «Известия», 30.10. 2002]

Таблица 1

	Нейтральный синоним	Переносное Значение	Оценочный компонент	Разговорный контекст / официальный контекст
Врубиться	*понять*	+	–	+/–
Зубрить	*заучивать*	–	+	+/–
Зазнаваться	–	–	+	+/ –
Баюкать	–	–	–	+/ –
Отмывать (деньги)	–	+	+	+/ +
Зачищать	–	+	–	+/+
Кидать	*обманывать*	+	+	+ / –
Крышевать	*защищать*	+	+	+/ –

У другого разговорного глагола, *зубрить*, есть нейтральный синоним *заучивать*, но значение *зубрить* не является метафорическим — иными словами, у этого глагола нет одного из пяти признаков.

Таким образом, носители языка и составители словарей считают разговорными глаголы, у которых есть не все, а хотя бы несколько из перечисленных признаков, но при этом степень разговорности глаголов — разная.

Можно представить результаты оценки степени разговорности нескольких глаголов по данным параметрам в таблице 1.

Прежде всего, из таблицы следует, что единственный обязательный признак, которым обладают все без исключения разговорные глаголы, — это использование в разговорном контексте. Именно с этим признаком связан разговорный характер глагола *баюкать,* самого «слабого» из приведенных в этой таблице: в силу ситуации, которую он описывает, он ассоциируется только с разговорными контекстами; никаких других признаков разговорности у него нет.

О решающей роли контекста в определении разговорности свидетельствуют и два других глагола: *отмывать* и *зачищать*. Хотя по происхождению оба эти глагола сленговые (*зачищать* — из военного жаргона, а *отмывать* — из криминального арго), сейчас оба они воспринимаются как нейтральные или очень близкие к нейтральным. Причина этого — в том, что наряду с разговорными контекстами, они широко употребляются в официальных.

Самый «сильный» с точки зрения разговорности — глагол *крышевать*, у которого есть все перечисленные признаки.

2. Стилистическая окраска отглагольных имен, производных от сленговых и разговорных глаголов

Словник Толкового словаря разговорной лексики позволяет опровергнуть мнение, согласно которому отглагольные имена имеют книжный характер, а их употребление характерно для письменной, в частности официальной речи: разговорные глаголы могут мотивировать как книжные, так и разговорные отглагольные имена. Рассмотрим, как связана стилистическая характеристика отглагольного имени с особенностями мотивирующего его глагола, характеризуя каждый глагол по параметрам, предложенным в предыдущей части (таблица 2).

Как следует из таблицы, от «слабых» разговорных глаголов *аукать, ахать, баюкать* образованы отглагольные имена книжной стилистической окраски *ауканье, аханье* и *баюканье*. Для стилистической окраски имен, образованных от более «сильных»

Таблица 2

	Нейтральный синоним	Метафор. значение	Оценочный компонент	Разг. / офиц. контекст	Отглагольное имя
Аукать	–	–	–	+ / –	*ауканье**
Ахать	–	–	–	+ / –	*аханье*
Бахвалиться	–	–	+	+ / –	*бахвальство*
Баюкать	–	–	–	+ / –	*баюканье*
Болтать 1 (самолет)	–	–	отр.	+ / –	*болтанка*
Болтать 2 (о чем-л.)	*говорить*	+	–	+ / –	*болтовня*
Брехать	*говорить*	+	отр.	+ / –	*брехня*
Вилять	–	+	отр.	+ / –	*виляние*

* Нейтральные и книжные отглагольные имена выделены заливкой.

	Нейтральный синоним	Метафор. значение	Оценочный компонент	Разг./ офиц. контекст	Отглагольное имя
Ворковать	–	+	–	+ / –	*воркотня / воркование*
Впарить	*продать*	+	отр.	+/ –	*впаривание*
Долбить	*учить*	+	отр.	+/ –	*долбежка*
Заварить	*затеять*	+	отр.	+ / –	*заваруха*
Заводить	*нервировать*	+	отр.	+ / –	*заводка*
Заделать	*законопатить*	–	–	+ / –	*заделка*
Зажимать	*подавлять*	+	отр.	+ / –	*зажим*
Заказать	*заплатить за*	+	отр.	+ / –	*заказ / заказуха*
Закатывать	*консервировать*	+	–	+ / –	*закатка*
Закачать (файл)	–	+	отр.	+ / –	*закачка*
заморачивать	*запутывать*	–	+	+ / –	*заморочки*
Замочить	*убить*	+	отр.	+ / –	*замочка***
Засветить	*заметить*	+	отр.	+ / –	*засветка*
Засыпать	–	+	отр.	+ / –	*засыпка*
Заточить	*подготовить, нацелить на*	+	–	+ / –	*заточка*
Зачистить	*очистить*	+	–	+ / +	*зачистка*
Крышевать	*защищать*	+	+	+ / –	*крышевание*
Отмывать	–	+	+	+ / +	*отмывание*

** Ср. : *Замочкой* в сортирах и массовыми зачистками, когда чаще погибают ни в чем не повинные люди, мира добиться невозможно (блоги).

разговорных глаголов, имеет значение характер суффикса: имена на *-ние / -нье* и на *-ство* получают книжную стилистическую окраску (*бахвальство, виляние*), а имена на *-ня* и *к(а)* — разговорную. Это наблюдение хорошая иллюстрируют два имени, производные от одного и тоже глагола *ворковать*: книжное *воркование* и разговорное *воркотня*.

Показательно, что характер суффикса определяет и стилистическую окраску отглагольных имен, мотивированных неразговорными (нейтральными) глаголами: отглагольное имя, образованное от нейтрального глагола с помощью разговорного суффикса, получает разговорную стилистическую окраску, например *оживляж* от *оживить / оживлять*, *визготня* от *визжать*. Между тем для отглагольных имен, мотивированных сленговыми глаголами, характер основы важнее, чем характер суффикса: они **всегда** имеют сленговую окраску, и даже «книжные» суффиксы не делают отглагольное имя книжным, ср. *впаривать — впаривание, крышевать — крышевание*. Тем более естественно, что сленговая окраска глагола наследуется отглагольными именами с «разговорными» суффиксами *-к(а)*, *-овк*, *-ов(о)*, *-ня*, *-ж* и нулевым суффиксом, ср.: *затачивать — заточка*, *кидать — кидаловка, кидалово, колготиться — колготня, залететь — залёт, гудеть — гудёж*[5].

Сфера функционирования разговорных и сленговых отглагольных имен — бытовая устная речь, публичная речь, публицистика, ср.:

(24) В цейтноте такая сложная *заваруха* могла окончиться как угодно! [Сергей Шипов. Счастливое число чемпиона — 14 (2004) // «64 — Шахматное обозрение», 2004.11.15]; Сложилась парадоксальная ситуация: согласно некоторым опросам, немногие россияне верят, что критики президента (например, коммунисты или яблочники), борясь за власть, способны устроить в стране какую-нибудь *заваруху*. [Павел Вощанов. Требуется тройник президента (2003) // «Новая газета», 2003.01.15].

Нужно сказать, правда, что начиная с 2004 г., использование разговорных и сленговых отглагольных имен в публицистике резко снижается.

[5] Не все разговорные и сленговые глаголы способны к номинализации. Вопрос о том, почему, при том, что от целого ряда разговорных глаголов возможно образование имен в значении результата (*залететь* ‘забеременеть случайно, вопреки желанию’ — *залет* ‘нежелательная беременность’, *заказать* ‘дать задание убить кого-л. за плату’- *заказ* ‘задание убить кого-л. за плату’) или процесса (*закатывать* ‘консервировать, делая края крышки герметичными с помощью специальной закаточной машинки’ — *закатка* ‘деятельность по консервированию и ее результат — герметично закрытая банка консервированных овощей или фруктов’ , *гудеть* ‘долго и много пить’ — *гудеж*), другие неспособны мотивировать отглагольные имена (например, *запасть на кого-л.* ‘влюбиться’- **запа’д*, *подсесть на что-л.* ‘пристраститься’ — **подсед*), представляет собой проблему, заслуживающую специального исследования.

Некоторые слова (например *замочка*, *засветка*) к 2005 г. перестают употребляться вообще; слово *зачистка* используется в 2005 г., только один раз по сравнению с 24 примерами употребления в 2004 г.

Между тем особенно высока частотность отглагольных имен в блогах — типе текстов, промежуточном между частным и публичным общением (в блогах пишут для себя и для друзей, но в то же время большая часть блогов открыта для прочтения всеми), устной и письменной речью (в блогах пишут, но в то же время всеми силами имитируют устную речь, так как настроены на получение ответных реплик прочитавших запись, т. е. на диалог). Слово *засветка*, о котором только что шла речь, появилось в блогах за последние 3 дня 6 раз; слово *зачистка* только за 7 часов 11 февраля 2010 года появилось на страницах блогов 19 раз! Приведу также несколько примеров из блогов на сайте радио «Эхо Москвы», иллюстрирующих, насколько широко используются разговорные и сленговые отглагольные имена в этом жанре текстов:

(25) Мужчины чаще всего думают о ногах, фигуре, внешности, но покажите мне женщину, которая сильно *западает* на ноги мужчины. <…> девственность — это самый дорогой товар женщины. И этот товар нельзя отдавать на дискотеках за *понюх* кокаина, его нельзя ориентировать на юношеские увлечения, на многочисленные влюбленности (Г. Стерлигов 9.02.2010).

(26) *Раскручивание* слогана (подзаголовок поста в блоге А.Илларионова. 9.02.2010).

Речь идет о *развале* существовавшей системы снабжения. (там же).

(27) Антисемитизм — прелестная вещь, многим нравится, но не универсальное алиби, не универсальная *отмазка* (Л. Радзиховский 11.02.2010).

Представляется что резкое сокращение использования сленговых и разговорных слов в СМИ при высокочастотном их использовании (возможно, возрастающем) в блогах отражает два разных явления: с одной стороны, появление внутренней цензуры в СМИ — в какой-то степени даже возвращение к стилистике советского времени (об этой же тенденции в связи с повышением частотности так называемых безагентивных предложений в языке современных газет говорится в работах [Кормилицына 2009] и [Сиротинина 2009: 10]), а с другой стороны — отказ от соблюдения норм и общее снижение стиля повседневного общения.

3. Семантика отглагольных имен

В статье [Mackenzie 2007] Л. Маккензи обращает внимание на то, что при номинализации возможен метонимический перенос, и в этом случае связь между отглагольным именем и мотивирующим его глаголом отсутствует. Речь идет о слове *impression*: так, в примере (8а) *impression* синонимично не предложению (8б), а словосочетанию (8в):

(28) а. His *impression* of Lacan ‘его образ / изображение Лакана’
б. He was *impressed* by Lacan ‘Лакан произвел на него впечатление’
в. His portrait of Lacan ‘Лакан в его изображении’, букв. ‘его портрет Лакана’[6].

На самом деле здесь имеет место не образование отглагольного имени с значением ‘образ’ от глагола *to impress* ‘произвести впечатление’, а полисемия отглагольного имени *impression* . В прямом значении ‘впечатление’ слово *impression* мотивировано конструкцией (9):

(29) The situation *impressed* me.

В свою очередь это значение отглагольного имени *impression* становится основой для переносного метонимического значения ‘образ / изображение’.

Слово *impression* в значении‘образ / изображение’ относится к классу имен способа (manner nominals), включающему такие отглагольные имена, как *интерпретация*, *исполнение*, *понимание*, *трактовка* и т. п. Общие свойства имен этого класса — их семантика и актантная структура детально описаны в работе [Падучева 2009].

В нашем материале не встречаются имена способа, однако статья Л. Маккензи важна для нас потому, что он обращает внимание на метонимические значения отглагольных существительных.

Наличие двух значений, прямого и переносного, встречается у отглагольных имен, имеющих разговорную или сленговую окраску, гораздо чаще, чем у номинализаций глаголов литературного языка, но этим различия между разговорными и литературными отглагольными именами не исчерпываются. В нашем материале встречаются примеры метонимической связи между значениями, аналогичные приведенному Л. Макензи, но есть и случаи метафорической связи.

---

[6] Это утверждение Макензи требует проверки: на самом деле, неясно, образованы ли оба значения (‘впечатление’ и ‘изображение’) от глагола *to impress* или же второе значение образовано от первого.

Слово *засыпка* в прямом значении — действие по глаголу *засыпать* в значении 1 ‘Забросать, заполнив доверху чем-л. сыпучим’ (МАС). От этого значения образовано два метонимических значения: ‘то, чем засыпают’ (*засыпка для супа*) и ‘то, чем засыпали’:

(30) Толстые брёвна стен, окошки-щели и крыша накатом с земляной *засыпкой* поверху. [Алексей Иванов. Сердце Пармы (2000)].

Кроме того, у слова *засыпка* есть метафорическое сленговое значение ‘провал на экзамене’ (в МАСе не отмечено), ср. *вопрос на засыпку*.

Отглагольное имя *долбежка* в прямом значении — действие по глаголу *долбить* ‘пробивать отверстие, делать углубление в чем-л. путем последовательных частых ударов какими-л. инструментами’ (МАС). У этого имени есть метафорическое разговорное значение ‘механическое заучивание, зубрежка’.

На самом деле, если между прямым и переносным метонимическим значением есть отношение производности, между прямым и метафорическим значением отглагольного имени оно отсутствует: прямое и метафорическое значения этих отглагольных имен — результаты номинализации мотивирующего глагола в разных значениях. Метафорическое значение отглагольного имени *долбежка* — результат номинализации глагола *долбить* в значении 4 по МАСу ‘механически повторяя, учить наизусть; зубрить’. Метафорическое значение отглагольного имени *засыпка* — номинализация глагола *засыпать* в другом, сленговом, значении ‘провалить на экзамене, задавая множество вопросов’.

Интересный пример, подтверждающий наше наблюдение о том, что прямое и переносное метафорическое значение разговорных отглагольных имен не связаны отношениями производности, но мотивированы разными значениями производящего глагола, представлен словом *заваруха*. В современной речи, в частности в публицистике, широкое распространение получило его употребление в разговорном значении ‘Сумятица, беспорядок, вызванные чем-л.’ (МАС)[7], которое мотивировано переносным просторечным значением глагола *заварить* ‘затеять, начать что-л. сложное, требующее хлопот, деятельности и т. п.’. Однако у слова *заваруха* было и литературное, теперь неактуальное значение — ‘каша, приготовленная определенным способом: заваренная кипятком’[8], ср.:

Заваруха-повалиха. В кипящую подслащенную воду всыпают просеянную пшеничную муку, проваривают до густоты манной каши. На смазанную жиром сковороду выкладывают горкой смесь, делают в середине углубление, наливают туда растопленный маргарин и запекают в духовке или в печи до румяной корочки. Подают с простоквашей (Рецепты старинных русских блюд. http://kuking.net).

Это значение результата — т. е. метонимическое значение отглагольного имени *заваруха*, мотивированное прямым значением глагола *заварить* с пропуском одного шага: образования отглагольного имени с значением действия.

В целом, характерную для отглагольных имен многозначность можно представить следующей схемой:

Глагол		Отглагольное имя
Значение 1 (прямое)	→	Значение 1 → значение 2 (метонимическое)
Значение n (разг. метафорическое)	→	значение 3 (разг. / сленг метафорическое)

Семантическая структура слова *зажим* подтверждает правомерность этой схемы. У этого отглагольного имени есть два литературных значения и одно разговорное. *Зажим* 1 в литературном языке — действие по глаголу *зажать* ‘обхватив, сдавить, стиснуть, защемить’ (МАС). Это значение мотивирует метонимически связанное с ним значение *зажим* 2 ‘приспособление для зажимания’. Между тем еще одно, разговорное метафорическое значение отглагольного имени *зажим* мотивировано другим, разговорным же значением глагола *зажать* ‘*перен. разг.* Стеснить в чем-л.; помешать свободному проявлению, развитию; подавить’ (МАС).

Производные значения отглагольных имен, мотивированные разными источниками — прямым значением отглагольного имени и переносным разговорным значением производящего глагола, обычно настолько далеко отстоят друг от друга, что можно говорить не о полисемии, а об омонимии этих имен.

Выводы

Итак, материал словника Толкового словаря разговорной лексики и примеры из Национального корпуса русского языка позволяют полностью развеять миф

[7] Не стоит думать, что это слово возникло, наряду с другими сленговыми словами, в перестроечный или постперестроечный период. Ровно в этом значении слова *заваруха* и *заварушка* зафиксировано в словаре Даля, ср.: Заваруха твер. свалка, ссора, свара, драка. Заварушка, то же, смятеньице, замешаньице, смуты, ссоры.

[8] вят. **заваруха** сиб. кисель, саламата, гуща; | новг. овсяная, ячная, пшенич. или ржаная кашка; з**аваруха**. крутая саламата, густая болтушка, горячий кисель, гуща; каша из ржаной муки, к завтраку, с молоком и с маслом, а в пост с медом.

о том, что отглагольные имена имеют книжный характер, а их употребление характерно для официально-делового стиля. На деле отглагольные имена, производные от разговорных и сленговых глаголов, широко употребляются в СМИ и блогах, представляющих собой жанр текстов, близких разговорной речи.

Стилистическая окраска отглагольных имен, мотивированных разговорными глаголами, определяется взаимодействием степени разговорности глагола и характером суффикса. Степень разговорности каждого глагола — производная от числа имеющихся у него признаков, характерных для типично разговорных глаголов. Отглагольные имена, мотивированные «слабыми» разговорными глаголами, имеют книжный характер. Стилистическая окраска отглагольных имен, производных от «сильных» разговорных глаголов определяется характером суффикса. Отглагольные имена, производные от сленговых глаголов, наследуют их сленговую окраску независимо от характера суффикса.

Для разговорных отглагольных имен характерно наличие метонимического значения, связанного отношением производности с прямым значением этого имени, и разговорного или сленгового метафорического значения, являющегося результатом номинализации метафорического значения мотивирующего глагола. Фактически, имеет место не полисемия отглагольного имени, а омонимия двух отглагольных имен, производных от разных значений мотивирующего глагола.

Я благодарю рецензента «Диалога» и Е. В. Падучеву за стимулирующие замечания и вопросы.

Литература

1. *Ермакова О. П., Земская Е. А., Розина Р. И.* Слова, с которыми мы все встречались. Толковый словарь Толковый словарь русского общего жаргона. М., 1999.
2. *Кормилицына М. А.* О некоторых характеристиках «постновояза» на страницах современной прессы // Проблемы речевой коммуникации. Вып. 9. Саратов, 2009.
3. *Курилова А. Д.* Толковый словарь разговорного русского языка. М., 2006.
4. *Пазельская А. Г. Аспектуальность и русские предикатные имена // Вопросы языкознания, М., 2003, №4, с.72–90.*
5. *Пазельская А. Г.* Валентные свойства русских отглагольных имён эмоций // Материалы международной конференции «Диалог 2005». М., 2005.
6. *Падучева Е. В.* Посессивы и имена способа действия // Компьютерная лингвистика и интеллектуальные технологии. Вып. 8 (15). По материалам международной конференции Диалог 2009.М., 2009.
7. *Розина Р. И.* Номинализации в разговорной речи // Материалы международной конференции «Диалог 2008». М.: РГГУ, 2008.
8. *Сиротинина О. Б.* Вероятное и возможное в судьбе русского языка (размышления на основе фактов его изменений в начале XXI века) // Проблемы речевой коммуникации. Вып. 9. Саратов, 2009.
9. *Толковый* словарь русского языка с включением сведений о происхождении слов // Под ред. Н. Ю. Шведовой. М., 2007.
10. *Химик В. В.* Большой словарь русской разговорной экспрессивной речи. СПб., 2004.
11. *Цзяхуа Ч.* Аспектуальные семантические компоненты в значении имен существительных в русском языке // Вопросы языкознания, М., 2007, № 1.
12. *Mackenzie J. L.* Double-possessive nominalizations in English // Functional perspectives on grammar and discourse. Papers in honour of Angela Downing. Amsterdam, 2007, p. 217–232.
13. *Rozwadowska B.* Towards a unified theory of nominalizations: external and internal eventualities. Wrocław: Wydawnictwo Uniwersitetu Wrocławskiego, 1997.

Идентификация авторства коротких текстов методами машинного обучения

Authorship identification of short texts with machine learning techniques

Романов А. С. (alexx.romanov@gmail.com),
Мещеряков Р. В. (mrv@keva.tusur.ru)

ГОУ ВПО «Томский государственный университет систем управления и радиоэлектроники», Томск

В статье рассматривается проблема идентификации авторства коротких текстов. Описан процесс формирования модели автора и алгоритм разбора текста. Приведено описание и результаты экспериментов по идентификации авторства коротких электронных сообщений в случае двух возможных альтернатив с помощью разновидностей искусственных нейронных сетей и аппарата опорных векторов.

1. Введение

Задача идентификации авторства коротких текстов возникает чаще, чем задача определения авторства текстов больших объемов и является в настоящее время актуальной проблемой. Это связано, прежде всего, с широким распространением программ для обмена сообщениями в сети Интернет (интернет-мессенджеров), возросшей роли электронной почты в деловой переписке, высокой популярности интернет-форумов и блогов. Пользователи имеют возможность отправлять сообщения без регистрации и указания какой-либо информации о себе, а регистрация сама по себе зачастую носит чисто символический характер. То же самое касается интернет-мессенджеров и электронной почты — регистрационные данные не позволяют однозначно идентифицировать личность собеседника, адрес отправителя можно легко изменить.

Идентификации авторства коротких текстов посвящено сравнительно небольшое количество работ. Стоит отметить, что авторам не известны подобные работы отечественных исследователей для русского языка. Судить о точности тех или иных методов по результатам исследования для английского и др. языков не корректно в силу особенностей строя каждого языка. В частности главной особенностью русского языка в сравнении с английским, для которого представлено большинство результатов, является его флективность, а, следовательно, и более сложное словообразование.

Эксперименты на корпусе электронных писем (всего 253 письма 4 авторов) в работе [1] дали итоговую максимальную точность 82,4 % при идентификации на основе 184 характеристик уровней символов и слов и признаков электронного письма (позиции цитат, доли слов приветствия, прощания, подписи к общей длине письма, количество вложений). При этом исследователи утверждают, что минимально необходимый объем письма для определения авторства составляет 200 слов, а для обучения модели достаточно 20 таких писем.

В работе [2] исследовался метод опорных векторов (SVM) на примере корпуса немецких газет «Berliner Zeitung» (2652 статьи, средний размер каждой из которых 200–300 слов). 2121 статья использовалась для обучения, тестирование проводилось на оставшихся 531 статьях, с последующей заменой тестовой и обучающей частей в соотношении 4 к 1. Средняя точность классификации по 7 авторам на основе словоформ (всего около 120 000 признаков) составила 99,7 %, а на основе сочетания грамматических классов, их биграмм и распределения длин слов в тексте — 99,2 %. Эксперименты показали, что выбор того или иного ядра для классификатора не играет существенной роли. Также в этой работе авторы сравнивают SVM с нейронными сетями и деревьями решений — машина опорных векторов и перцептрон показали сравнимые результаты (100 % и 93,3 % соответственно), тогда как деревья решений с задачей успешно справиться не смогли (точность 22,7 %).

В работе [3] для идентификации автора электронных писем применялся метод k ближайших соседей, точность при этом в среднем составляла 80 %.

Эксперт в области криминалистической лингвистики Кэрол Часки в работе [4] для идентификации авторства коротких текстов из области крими-

налистики применяла линейный дискриминантный анализ и текстовые аномалии на всех лингвистических уровнях. Точность идентификации в зависимости от используемого типа ошибок, допущенных автором, в её работе колебался от 65 % до 92 %.

Аббаси, исследовавший авторство сообщений на английском и арабском языке с интернет-форума экстремисткой группы, утверждает, что метод опорных векторов справляется с этой задачей лучше, чем деревья принятия решений, искусственные нейронные сети и линейный дискриминантный анализ, которые в свою очередь превосходят по точности методы неконтролируемого обучения такие как метод главных компонент. Точность классификации с помощью SVM в его работе [5] по 5 авторам для английского языка составляет 97 %, для арабского языка — 94,83 %. Аналогичные исследования для онлайн сообщений на китайском языке проводились в работе [6]. Точность классификации с помощью SVM составила 88,33 %, с помощью нейронных сетей — 83,05 %.

Целью данной работы является проверка способности современных машинных методов обучения, таких как искусственные нейронные сети и машина опорных векторов, идентифицировать автора короткого электронного сообщения на русском языке.

2. Модель автора и текста

Проблему идентификации автора текста при ограниченном наборе альтернатив сформулируем следующем образом. Имеется множество текстов $T = \{t_1,...,t_k\}$ и множество авторов $A = \{a_1,...,a_l\}$. Для некоторого подмножества текстов $T' = \{t_1,...,t_m\} \subseteq T$ авторы известны, т. е. существует множество пар «текст-автор» $D = \{(t_i, a_j)\}_{i=1}^{m}$. Необходимо установить, кто из множества A является истинным автором остальных текстов (анонимных или спорных) $T'' = \{t_{m+1},...,t_k\} \subseteq T$.

В данной постановке задачу идентификации автора можно рассматривать как задачу классификации с несколькими классами [7]. В этом случае множество *A* составляет множество предопределенных классов и их меток, *D* — обучающие примеры, а множество T'' — классифицируемые объекты. Целью является построение классификатора, решающего данную задачу, т. е. нахождение некоторой целевой функции $F : T \times A \to [-1,1]$, относящей произвольный текст множества *T* к его истинному автору. Значения функции интерпретируется как степень принадлежности объекта классу: 1 соответствует полностью положительному решению, −1 — отрицательному.

Для решения задачи можно использовать любой из разработанных на данный момент алгоритмов классификации. IDEF0 модель процесса формирования модели автора показана на рис. 1.

Для определения отличий стилей авторов предлагается следующая последовательность действий:

1. Разбиение имеющегося множества текстов на две группы. Первая используется для обучения модели классификатора. Вторая — для проверки точности идентификации автора с помощью обученной модели.
2. Формирование модели текста путем выбора модели представления текстовой информации и выделения определенных информативных групп характеристик текста. Отличия в стилях авторов характеризуются главным образом употреблением и частотой встречаемости определенных признаков в тексте — вектором $(x_1, x_2,..., x_n)$.
3. Приведение значений признаков в единый диапазон с помощью операций нормирования и шкалирования.
4. Корректировка параметров классификатора, позволяющих обеспечить высокую разделяющую способность исследуемых авторов путем обучения классификатора на нормированных векторах признаков группы обучающих текстов и проверке точности обученного классификатора на векторах признаков тестовой группы текстов. Первоначальное обучение классификатора происходит с параметрами по умолчанию или при заданных параметрах.
5. Изменение перечня групп характеристик и/или признаков, составляющих группу, в случае если изменением параметров классификатора достичь требуемой точности не удается.

Итогом является обученный классификатор, веса связей которого настроены таким образом, чтобы он был способен разделить стили авторов, на текстах которых проводилось обучение, при подаче на его входы подобранного набора признаков.

Таким образом, конечная модель помимо информативности признаков текста, учитывающихся в статистических методах идентификации авторства, учитывает влияние общей способности классификатора к разделению данных и его точность.

В данной работе были выбраны два инструмента — искусственные нейронные сети (многослойные перцептрон и сети каскадной корреляции) и аппарат опорных векторов. Эксперименты отечественных и зарубежных исследователей показывают, что на сегодняшний день эти два инструмента, при должной настройке и выборе входных параме-

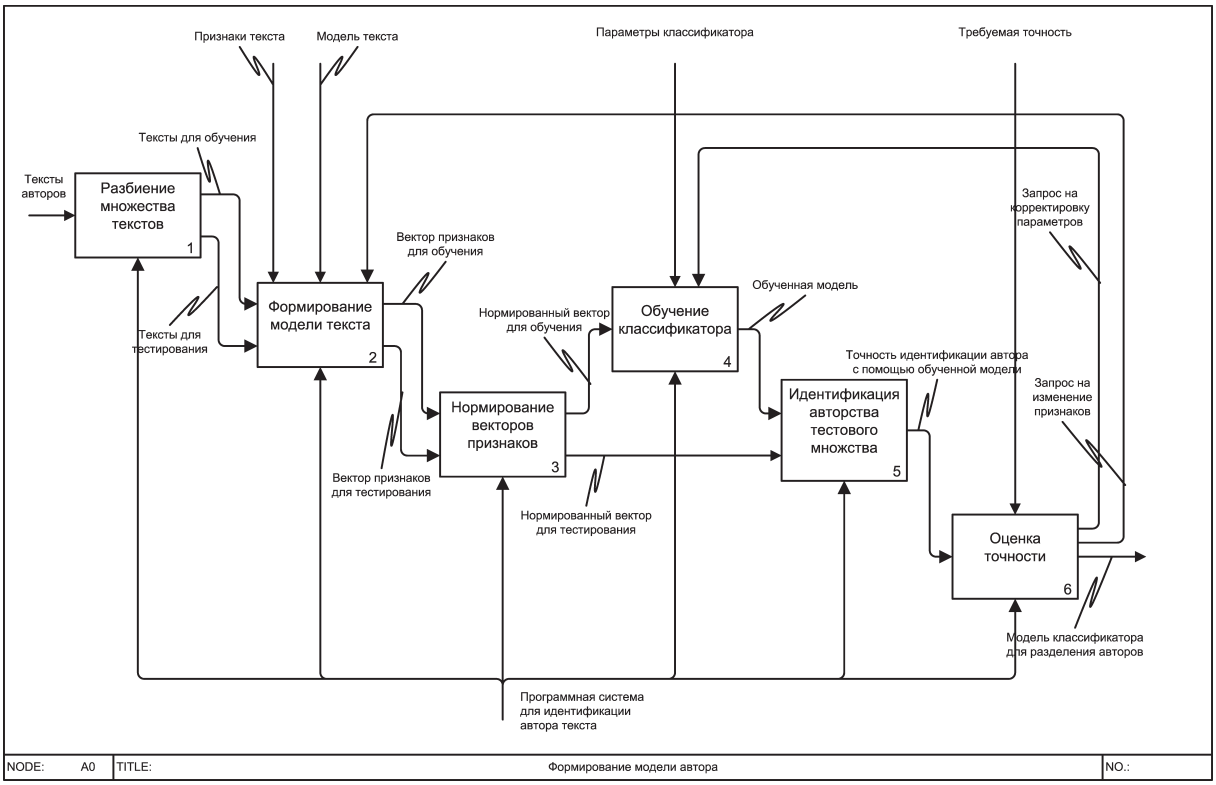


Рис. 1. IDEF0 модель процесса формирования модели автора

тров, являются лучшими в своем классе. Но каждому из них присущи свои достоинства и недостатки. Так, например, многослойный перцептрон (MLP) [8] долго обучается и не может работать с большим признаковым пространством. Временные издержки на подбор топологи сети и обучение можно сократить, применяя алгоритмы построения сетей с оптимальной топологией, к которым относятся сети каскадной корреляции (CCN) [9,10], однако точность классификации при этом может существенно снизиться. Метод опорных векторов [11, 12] лишен всех этих недостатков, однако слишком чувствителен к шумам во входных данных. Поэтому необходимо детально изучить возможность применения этих инструментов к задаче идентификации авторства, сравнить их производительность и эффективность.

В модели используется векторное представление текста, когда каждый текст представлен точкой в N-мерном пространстве. Элементами вектора могут быть характеристики уровней символов, слов, предложений, структурные признаки текста.

3. Алгоритм разбора текста

Особенностью коротких электронных сообщений является использование авторами эмотиконов («смайликов»), отсутствие пунктуации, намеренное искажение слов и т. д. Учет этой информации при определении авторства требует использования модифицированных алгоритмов разбора текста.

Специфика алгоритмов разбора текста состоит в том, что в процессе их работы происходит посимвольный или пословный анализ всего текста, в результате которого проверяется, удовлетворяет ли данная последовательность символов или слов определенной группе эвристических правил. Таким образом, для вынесения итогового решения, проверяется несколько промежуточных условий. Текущее состояние решения и очередной символ, поданный на вход алгоритма, определяют следующее состояние. Очевидно, что наилучшим техническим решением при реализации данной группы алгоритмов является использование конечных автоматов, несомненным преимуществом которых является возможность расширения функциональности алгоритма за счет добавления новых состояний.

Диаграмма состояний графа для определения границ предложений в коротких электронных сообщениях на рис. 2.

Началом предложения считается первый печатный символ текста. Концом предложения помимо последнего символа сообщения считается точка, вопросительный или восклицательный знак или их группа, а также любой эмотикон. Эмотиконы в большинстве случаев выражают законченность мысли и служат для придания написанным словам дополнительной эмоциональной окраски, тогда как в середине предложения употребляются редко. Также они используются в начале сообщения,

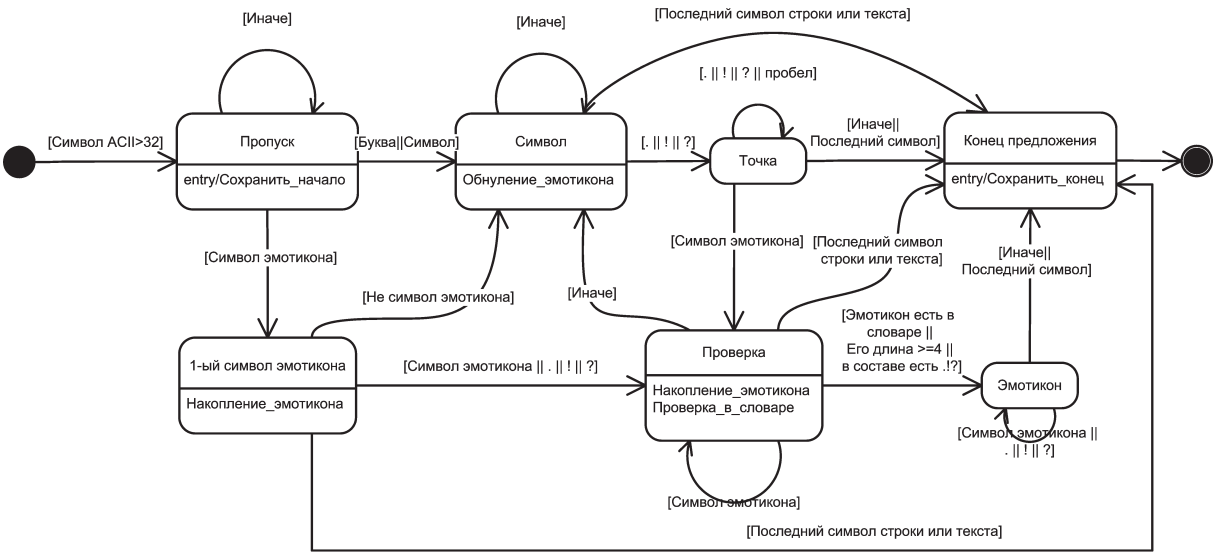


Рис. 2. Алгоритм определения границ предложений в коротких сообщениях

чтобы выразить эмоции по отношению, например, к предыдущей фразе собеседника — в этом случае алгоритм не выделяет эмотикон как отдельное предложение, а включает его в состав первого предложения сообщения.

Разделителями слов выступают все символы, не относящиеся к русскому и латинскому алфавиту, а также цифры. Один или несколько последовательно идущих символов «.», «!», «?» рассматриваются как отдельная группа символов конца предложения. В алгоритме также предусмотрен ряд проверок символа «-», позволяющих разделять случаи, когда он является переносом, частью составных слов, написанных через дефис и т. д.

4. Экспериментальная часть

В качества корпуса для исследования были взяты сообщения 10 авторов, собранные на Интернет-форуме http://forum.tomsk.ru. Информация о корпусе представлена в табл. 1 (длина текста в символах).

Все сообщения были разбиты на тексты длиной 100 символов (сообщения меньшего объема использовались целиком). Для обучения классификаторов использовалось по 50 таких текстов, тестирование проводилось на 25 сообщениях каждого автора. После этого обучающая и тестовая части менялись и опыты повторялись, пока каждый из текстов не был использован в тестовой части. Всего было проведено около 3000 таких экспериментов с различными комбинациями авторов и сообщений, что обеспечило необходимое покрытие сочетаний букв и слов русского языка.

В работе был рассмотрен самый простой случай, когда для спорного текста существует два предполагаемых автора. Данный частный случай был рассмотрен специально, чтобы раскрыть все потенциальные возможности метода опорных векторов, который по умолчанию является бинарным классификатором и поставить его в равные условия с двумя другими методами применительно к русскому языку.

Таблица 1. Корпус текстов для исследований

Автор	Количество текстов	Средняя длина текста	Минимальная длина текста	Максимальная длина текста
1	126	310,8	80	1368
2	102	438,2	83	2100
3	180	243,6	69	869
4	145	308,4	81	1665
5	141	238	72	928
6	146	261,1	83	1264
7	186	425,2	89	1649
8	116	292,4	68	1130
9	140	329	94	1957
10	157	539,8	95	3407
Все авторы	143,9	339,1	68	3407

Исследования проводились с использованием программной системы для идентификации автора письменной речи «Авторовед» [13].

Параметры обучения нейронных сетей архитектуры многослойный перцептрон были выбраны следующие:

- алгоритм обучения — обратного распространения ошибки;
- функция активации скрытых слоев — сигмоид;
- функция активации выходного слоя — сигмоид;
- скорость обучения — 0,7;
- момент — 0,0;
- количество скрытых слоев — 1;
- количество нейронов в скрытом слое — 10;
- максимальное количество эпох обучения — 20 000;
- допустимый уровень ошибки — 0,00 001.

Параметры обучения нейронных сетей каскадных корреляций были выбраны следующие:

- алгоритм обучения — быстрого распространения ошибки;
- функция активации скрытых слоев — сигмоид;
- функция активации выходного слоя — сигмоид;
- максимальное количество нейронов, которое можно добавить — 20;
- допустимый уровень ошибки — 0,00 001.

Параметры обучения метода опорных векторов были выбраны следующие:

- алгоритм обучения — метод последовательной оптимизации;
- ядро — линейное;
- параметр регуляризации $C = 1$;
- допустимый уровень ошибки — 0,00 001.

В качестве характеристик использовались признаки текста, показавшие наилучшие результаты на литературных текстах большего объема в ранней работе авторв [7]:

- частоты наиболее частых слов русского языка (согласно словарю Шарова [14]);
- частоты наиболее частых триграмм русского языка;
- частоты униграмм символов;

К признакам также добавлены частоты отдельных знаков препинания, составные знаки препинания (многоточие, «!?!» и т. п.) и эмотиконы.

Итоговый вектор признаков содержит 1024 компоненты. Данный вектор характеризует строй русского языка как лексико-зависимый, имеющий четко выраженную структуру, использующий базу наиболее употребимых слов и трехбуквенных сочетаний символов.

Результаты исследований представлены в табл. 2.

Таблица 2. Результаты экспериментов

	MLP	**CCN**	**SVM**
Точность классификации	0,62±0,01	0,70±0,17	0,69±0,12
Время обучения, с.	503,1	206,15	0,6

Из таблицы видно, что наименее точным в данном случае является многослойный перцептрон. Также этот метод классификации оказался наиболее затратным по времени. Применение сетей каскадных корреляций позволяет существенно уменьшить временные затраты на обучение модели и повысить точность идентификации. Дополнительные исследования по подбору архитектуры многослойного перцептрона, возможно, позволили бы повысить его точность до уровня сетей каскадных корреляций, однако это потребовало бы существенных временных затрат. Сравнимую с сетями каскадных корреляций точность показывает классификатор на основе машины опорных векторов. Небольшие потери в точности компенсируются высокой скоростью обучения моделей.

Точность идентификации выше 0,5 позволяет позволяет сделать вывод о принципиальной возможности идентификации авторства коротких текстов именно для русского языка. Несомненным положительным результатом можно признать учет лексической и синтаксической специфики русского языка, на основе которой удалось определить характеристики короткого сообщения, имеющие преимущественное значение для использования в методиках определения автора короткого сообщения и отличающиеся от других языков.

Дальнейшие исследования авторов по данной тематике будут связаны с применением ансамблей классификаторов для идентификации авторства и поиском статистически устойчивых характеристик на малых текстовых фрагментах.

Работа поддержана грантом ФСРМПНТ.

Литература

1. *Corney M., Anderson A., Mohay G., De Vel O.* Identifying the Authors of Suspect E-mail [Электронный ресурс] // Computers and Security, 2001. — Режим доступа: http://eprints.qut.edu.au/8021/1/CompSecurityPaper.pdf, свободный.
2. *Diederich J., Kindermann J., Leopold E., Paass G.* 2003. Authorship attribution with support vector machines. Appl. Intell. 19, pp. 109–123.
3. *Calix K., Connors M., Levy D.* Stylometry for E-mail Author Identification and Authentication // Proceedings of CSIS Research Day, Pace University, May 2008. — Режим доступа: http://csis.pace.edu/~ctappert/srd2008/c2.pdf, свободный.
4. *Chaski C. E.* Who's at the keyboard: Authorship attribution in digital evidence investigations // International Journal of Digital Evidence, vol. 4, no. 1, p. n/a, Electronic-only journal: http://www.ijde.org, accessed May 31, 2007, 2005.
5. *Abbasi, Chen H.* Applying authorship analysis to extremist-group web forum messages // IEEE Intelligent Systems, vol. 20, no. 6, pp. 67–75, 2005.
6. *Zheng R., Li J., Chen H., Huang Z.* A framework for authorship identification of online messages: Writing-style features and classification techniques // Journal of the American Society for Information Science and Technology, vol. 57, no. 3, pp. 378–393, 2006.
7. *Романов А. С., Мещеряков Р. В.* Идентификация автора текста с помощью аппарата опорных векторов // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2009» (Бекасово, 27–31 мая 2009 г.). Вып. 8 (15). — М.: РГГУ, 2009. — С. 432–437.
8. *Хайкен С.* Нейронные сети: полный курс / Саймон Хайкен. — 2-е изд. — М. : Вильямс, 2006. — 1104 с.
9. *Fahlman S. E., Lebiere C.* The cascade-correlation learning architecture. Tech. Rep. CMU-CS-90–100, School of Computer Science, Carnegie Mellon University, August 1991.
10. *Hoefeld M., Fahlman S. E.* Learning with limited numerical precision using the cascade-correlation algorithm. Tech. Rep. CMU-CS-91-130, School of Computer Science, Carnegie Mellon University, 1991.
11. *Vapnik V. N.* Statistical Learning Theory // Wiley, New York, 1998. — 732 pages.
12. *Vapnik V. N.* The nature of statistical learning theory // Springer-Verlag, New York, 2000. — 332 pages.
13. *Романов А. С.* Структура программного комплекса для исследования подходов к идентификации авторства текстов // Доклады Томского государственного университета систем управления и радиоэлектроники. Томск: Изд-во ТУСУР, 2008. Ч.1. №2(18). С. 106–109.
14. *Шаров С. А.* Частотный словарь [Электронный ресурс]. — Режим доступа: http://www.artint.ru/projects/frqlist.asp, свободный. — Загл. с экрана.

Опыт автоматизированного пополнения онтологий с использованием машиночитаемых словарей

The approach to ontology learning from machine-readable dictionaries

Рубашкин В. Ш. (vrubashkin@yandex.ru),
Бочаров В. В. (victor.bocharov@gmail.com),
Пивоварова Л. М. (lidia.pivovarova@gmail.com),
Чуприн Б. Ю. (boris@vr4591.spb.edu)

Санкт-Петербургский государственный университет

В докладе представлено текущее состояние разрабатываемой авторами технологии пополнения онтологии на основе лингвистического и семантического анализа текстов определений слов (терминов) в энциклопедических и толковых русскоязычных словарях.

1. Положение дел и постановка задачи

Проблема массового пополнения онтологий — с расчетом на достаточно полное покрытие текстов хотя бы в рамках ограниченной предметной области — характеризуется обычно как проблема «бутылочного горлышка» (*bottle neck*). Эта характеристика отражает факт высокой трудоемкости «ручного» (лучше сказать — экспертного) ввода концептов в онтологию и отсутствие широко признанных технологий, обеспечивающих хотя бы частичную автоматизацию процедур ввода. Следует подчеркнуть, что в этом контексте речь идет не о *формировании*, а именно о *массовом пополнении* онтологий. Первоначальное формирование как специализированной, так и в особенности универсальной онтологии предполагает построение системного ядра, состоящего из "высокоуровневых понятий» (*Top-Level Ontology*). Построение такого ядра в любом случае есть дело высококвалифицированного специалиста. В докладе представлен опыт разработки и использования человеко-машинной технологии ввода, опирающейся на использование традиционных словарных ресурсов и ориентированной в основном на пополнение уже сформированного ядра онтологии концептами, именующими объекты («предметные имена»).

В рамках общего дискурса *Ontology Learning* обсуждаются два пути автоматизации пополнения онтологий[1]:

1) на основе анализа дистрибутивных характеристик лексики в корпусе текстов;
2) с использованием традиционной лексикографической информации — энциклопедических и толковых словарей (в англоязычной литературе — *машиночитаемые словари — MRD*).

Изучение результатов, полученных на том и другом направлении, а также наш собст томатического построения таксономий и извлечения других семантически значимых отношений была доведена до технологической реализации, была представлена еще в 1985 г. ([2]). В 1993 г. N. Ide и J. Veronis уже подводят, так сказать, промежуточные итоги ([3]), скептически оценивая как полученные результаты, так и вообще перспективы полностью автоматической «генерации больших баз знаний» из MRD. Их основная аргументация связана с семантическим качеством самих словарных определений. Однако это не остановило дальнейших попыток, и в последнее десятилетие работы в этом направлении развивались широким фронтом — главным образом, на материале англоязычных словарей (ср., напр., [4], [5], [6]).

Другой наш предварительный вывод состоит в том, что при современном уровне развития технологий автоматического понимания текста не следует рассчитывать на то, что распознавание востребованных онтологией словарных характеристик может быть обеспечено исключительно алгоритмическими средствами; оперативное участие администратора онтологии в практически полезной техно-

[1] Подробный обзор сложившихся подходов можно найти в монографии [1].

логии такого рода необходимо.[2] Наш общий подход при этом состоит в том, что средства автоматизации должны обеспечить предварительную структуризацию текстов определений, поддерживать отбор релевантной задаче лексики предметной / проблемной области в машиночитаемом словаре и осуществлять предварительную онтологическую квалификацию содержащегося в словаре термина (не всегда точную и не всегда правильную); роль специалиста, участвующего в этой работе, сводится к тому, что он оценивает предлагаемые системой решения и либо просто подтверждает их, либо вносит необходимые коррективы, т. е., предлагается ориентация на человеко-машинную технологию и на создание обучаемой диалоговой среды пополнения онтологии. Возможности графического редактирования, ставшие стандартом де-факто для современных онторедакторов, существенно облегчают и упрощают эту часть работы.

Основная идея, положенная в основу рассматриваемой технологии, состоит в следующем ([2], [8]). Определения и толкования в энциклопедических и толковых словарях представляют собой частично структурированные тексты. Самое существенное в этом плане то, что определение / толкование в подавляющем большинстве случаев соответствует стандартной логической схеме — «*род — видовые отличия*». Номинальное определение термина в энциклопедическом словаре почти всегда локализовано в первом предложении словарной статьи; в толковых словарях оно вообще, как правило, состоит из одной короткой фразы. При этом логически самая важная часть определения, указывающая родовое понятие, с помощью которого описывается определяемый термин, почти всегда грамматически акцентирована — представлена именной группой с главным словом — существительным в именительном падеже. Несколько примеров. Толковый словарь Ожегова: *МУШКЕТЕР — солдат, вооруженный мушкетом; МУШКЕТ — старинное ружье крупного калибра с фитильным замком.* Энциклопедический словарь [9]: *МАГНЕТРОН — электровакуумный прибор, мощный генератор электромагнитных волн сантиметрового диапазона. Принцип действия магнетрона основан на торможении электронов в скрещенных электрических и магнитных полях. Используется главным образом в устройствах радиолокации, а также в нагревательных установках сверхвысокой частоты.*

Следовательно, первоочередным объектом анализа должна быть указанная именная группа и ее главное слово («опорное слово») прежде всего. Исключения обозримы и обрабатываются спецалгоритмами, позволяющими в большинстве случаев найти «правильное» опорное слово.[3]

Помимо этого в тексте определения, как правило, присутствуют типовые элементы описания, указывающие функциональность, принцип действия, сферу использования, структурно-морфологические характеристики и свойства определяемого объекта/явления, а также, возможно, указаны отношения определяемого с другими понятиями, в частности, с понятиями, уже представленными в онтологии и/или словаре. Конечная цель состоит в том, чтобы они были распознаны и отражены как онтологические характеристики соответствующих концептов.

Понятно, что в такой постановке рассматриваемая задача, сводится к двум подлежащим алгоритмизации процедурам. Первая — структурирование текста определений (*подсистема анализа словарных определений*), вторая — сопоставление таковых с лексическим представлением концептов в онтологии (далее будем именовать ее *подсистема энциклопедического импорта*). При этом вторая процедура может быть организована одним из двух альтернативных способов. Либо для каждого определяемого термина словаря проводится поиск возможных релевантных родовых концептов — для последующего включения термина в таксономию, либо, наоборот, можно указать иерархически организованный фрагмент онтологии, требующий пополнения, и выбирать в множестве структурированных определений те, которые соответствуют указанному фрагменту, уточняя далее тем или иным способом место выбранной группы термина в этом фрагменте онтологии и, если нужно, дополнительные словарные характеристики отдельных терминов. Ясно, что первый подход эффективнее в ситуации, когда соответствующий раздел онтологии уже в первом приближении сформирован, и речь идет о его дополнениях и уточнениях. Второй подход эффективнее может применяться в той ситуации, когда нужно осуществить начальное наполнение определенного раздела.

Имеется в виду, что названные две процедуры работают автономно, интерфейс между ними определяется передаваемой структурой данных — в форме таблиц РБД. Ясно также, что оптимальным

---

[2] Ср. мнение авторитетных японских специалистов T. Morita и T. Yamaguchi, формулируемое в обоснование принятого ими подхода в проекте DODDLE [7]: "*Regarding domain ontology development support, many tools have been done with knowledge engineering, natural language processing and data mining techniques to make possible automatic domain ontology construction from existing information resources… However, as the techniques are not yet mature to achieve the task and domain ontology structure depends on the aspects from human experts (users), full automatic process does not go well with the task. Instead of developing full automatic environment, it is more important to provide refined semi-automatic environment with integrated facilities to construct practical domain ontologies*".

[3] Обзор типовых случаев невыполнения общего правила дан в [8]. Разумеется, предлагаемым способом не могут быть обработаны «недоброкачественные» определения, — скажем такого рода: *Боразол — B3H6N3* ; им присваивается в позиции опорного слова помета "*unknown*".

решением с точки зрения организации инструментальных средств здесь является погружение средств импорта в среду онторедактора, с тем, чтобы максимально использовать его функциональность — что и осуществлено в настоящей работе.

2. Онтология и онторедактор.

Базовой структурой любой онтологии является ***таксономия объектов***. В разработанной и используемой нами онтологии (онтология *InTez*) [10] таксономия представлена в форме ***дерева признаков*** [11]. Это структура, которая позволяет наиболее естественным образом отображать в онтологии связи типа *признак — значения признака; применимость признака к классу объектов*; соответственно, вычислять полный набор объемных отношений (*включение, совместимость, несовместимость*) между концептами — классами. Помимо объемных отношений, в онтологии представлены нетаксономические («ассоциативные») отношения — как универсальные (*часть — целое, объект — место, объект — функция*), так и специализированные (*страна — столица, руководитель — организация* и т. д.). Для данной задачи существенно, что онтология является *интерпретируемой*, т. е. поддерживает связь с лексикой естественного языка фиксируя двустороннее соответствие вида

КОНЦЕПТЫ ←→ СЛОВА,

достаточное для фиксации отношений омонимии и синонимии в естественном языке, с одной стороны, и учета многообразия лексической реализации концептов, с другой. Функционально это соответствие представляет собой одновременно толковый словарь и словарь синонимов — разумеется, в пределах наличного концептуального и лексического состава онтологии.

Операционально место концепта в данной онтологии определяется либо указанием классификационного признака, представляющегося основание, по которому выделен именуемый простой класс (*водный транспорт, воздушный транспорт, сухопутный транспорт → по среде перемещения*). либо — в случае многоаспектного определения класса — формальным толкованием, в типовом случае представляющим конъюнкцию простых классов (как в примере с термином *магнетрон: электровакуумный_прибор & генератор_электромагнитных_волн*). В последнем случае эквивалентным построению формального толкования является указание множественного наследования в таксономии. Возможны и другие схемы формальных толкований.

С точки зрения рассматриваемой задачи существенно то, что описание концептов в онтологии не требует столь подробной характеризации определяемого концепта, какая ему дается в энциклопедических словарях. Так, в приведенном выше определении термина *магнетрон* онтологически значимым можно считать лишь первое предложение определения. Второе предложение для формального описания концепта в онтологии избыточно. Поэтому перед средствами лингво-семантического анализа определений не стоит задача полной формализации всего текста. Скорее речь идет о методах целенаправленного поиска и распознавания онтологически значимых элементов.

Оптимальное распределение работы между программными процедурами и администратором онтологии в значительной степени обусловлено **функциональностью онторедактора**. В онторедакторе, поддерживающем работу с онтологией *InTez* [12], имеются, в частности, стандартные средства навигации, графического редактирования таксономии, поиска концептов по вариантам лексического представления, а также средства добавления связей между концептами. Частью онторедактора является также машина ограниченного вывода, являющаяся важным компонентом для подсистемы логического контроля при вводе. Это создает достаточно комфортную среду для быстрого постредактирования добавленных автоматически концептов — если в этом возникает необходимость.

3. Анализ словарных определений

Представляет собой последовательность следующих шагов.[4]

1) лексикографический парсинг словарных статей;
2) выделение опорного слова;
3) уточнение (при необходимости) опорного слова;
4) уточнение формулировки родового понятия присоединением зависимых и других связанных по смыслу с опорным слов.
5) построение формального толкования определяемого термина на основе концептуального содержания, представленного в текущей версии онтологии.

На данном этапе проекта реализованы в достаточно полном объеме п. п. 1, 2 и 3 и в очень ограниченных пределах — п. 5. Функциональность, соответствующая п. 4, частично реализована в подсистеме энциклопедического импорта.

Лексикографический парсинг предназначен для того, чтобы подготовить текст словарной статьи для синтаксического анализа; он включает удаление всех словарных помет, восстановление сокращений, «расклейку» множественных определений.

---

[4] Более подробно см. в [13].

Можно считать, что пункт 2 при доступных сегодня средствах анализа представляет собой чисто технологическую процедуру и состоит из следующих действий:

- получение грамматических описаний и лемматизация лексики первого предложения словарного определения;
- поиск и извлечение опорного слова (выбирается первое существительное, имеющее признаки: *существительное*, *именительный падеж*);
- частичная синтаксическая разметка первого предложения (анализ контактных связей).

Омонимия грамматических описаний сохраняется; ее частичное разрешение происходит по синтаксическому контексту — по результатам, полученным на этапе синтаксической разметки. Сохраняются также все полученные варианты синтаксической разметки (глобальная синтаксическая омонимия при анализе не рассматривается).

Используемые средства:

- морфологический анализатор АОТ [14];
- синтаксический анализатор АОТ, реализующий алгоритм GLR парсинга, со специально разработанной упрощенной грамматикой, предназначенной для анализа онтологически значимых фрагментов текстов определений;

На выходе процедуры анализа порождаются две таблицы РБД (СУБД *MS SQL Server,* используемая в качестве операционной среды редактора онтологии). Таблица «Термины» содержит термин, опорное слово, определение / толкование (множественные толкования «расклеены» на предыдущем шаге обработки), и дополнительную рабочую информацию. Таблица «*Слова*», связываемая с таблицей «*Термины*», содержит слова первого предложения определения термина, их леммы и части речи, а также элементы синтаксической разметки (*ссылка на синтаксического хозяина, вид связи*). Указанная пара таблиц далее модифицируется процедурами, соответствующими п. 3, после чего передается в подсистему энциклопедического импорта.

Процедура уточнения опорного слова (п. 3) в основном состоит (а) в удалении избыточных слов, с одновременным отысканием «истинного» опорного слова (слова с общим значением '*именование*' и '*принадлежность к классу*'); (б) в логической интерпретации опорных слов с общим значением '*отношение*' (*часть-целое*, *локализация, назначение,…*) и поиском второго объекта, с которым должно быть установлено распознанное отношение.

4. Энциклопедический импорт

Для организации взаимодействия с администратором онтологии в пользовательский интерфейс онторедактора добавлена вкладка *ТерминыСловаря*, отображающая содержимое таблицы *Термины.* Вкладка становится доступна в режиме «Энциклопедический импорт».

Процедура добавления группы терминов состоит из следующих действий.

1. Формирование группы добавляемых терминов: (а) по указанному опорному слову; (б) по лексике куста. Во втором случае администратор должен предварительно указать в онтологии (в дереве признаков) вершину куста, лексика которого должна учитываться при формировании энциклопедической выборки. Все имена концептов (включая синонимы), как непосредственно входящих в куст, так и связанных с концептами куста отношением «*НИЖЕ*» (*конкретизация*), — если они совпали с опорными словами, — включаются в список опорных слов, по которым формируется выборка. Полученная выборка может дополнительно уточняться (ограничиваться) двумя способами — одновременно обоими или любым из них по отдельности:
   1.1. По лексическому составу текста определения. Отбираются только термины, в тексте определения которых найдена заданная комбинация слов (возможно с усечением).
   1.2. По синтаксически зависимым от опорного словам определения. В этом случае строится и предъявляется администратору онтологии частотный словарь всех идентифицированных синтаксической разметкой слов (используются леммы), непосредственно зависимых от выбранного опорного слова (списка опорных слов) Администратору предоставляется возможность указать слова, уточняющие смысл опорного слова, объединив их связкой ***And*** или ***Or***.

   Кроме того, администратор имеет возможность «вручную» исключить из выборки или присоединить к выборке любые термины, представленные в одноименной таблице, устанавливая или снимая в соответствующих записях помету «Выбрано».
2. Автоматический ввод терминов, место которых в онтологии может быть определено самим алгоритмом. По каждому из идентифицированных системой терминов выборки, сформированной в соответствии с п. 1, администратору предъявляется предлагаемое алгоритмом решение, в отношении которого он может:
   (а) согласиться — и тогда система вводит его с определенными системой характеристиками;
   (б) отвергнуть — и тогда термин остается в общей выборке релевантных теме энциклопедических терминов, либо удаляется из выборки;

(в) отредактировать словарное описание по своему усмотрению и завершить редактирование командой ввода.

3. Для остающейся части выборки — выбор администратором базового концепта онтологии. Выбор производится в дереве признаков; может быть выбран концепт — простой класс, либо наименование классификационного признака.
4. Выбор способа добавления.
   4.1. Если администратором в качестве базового указано *наименование классификационного признака*, доступное действие — «*Присоединить к признаку*». В этом случае все термины выборки определяются как несовместимые подклассы, выделяемые по основанию, определяемому выбранным признаком.
   4.2. Если администратором в качестве базового указан *простой класс*, доступные действия:
      4.2.1. «*Добавить как синоним*» — термин добавляется как синоним указанного концепта.
      4.2.2. «*Добавить к новому классификационному признаку*» — система создает классификационный признак, определяющий новое основание деления указанного класса, запрашивая у администратора способ его именования. Все термины выборки подчиняются вновь созданному признаку, образуя разбиение исходного класса на подклассы.
      4.2.3. «*Добавить к новому списочному признаку*» — то же, что в п. 4.2.2. Но при последующей логической обработке вновь образованные подклассы не считаются объемно несовместимыми. Дальнейшая детализация таких подклассов в онтологии не допускается.
5. Если необходимо выполняется — графическое, либо символьное редактирование концептов добавленной группы (напр., перетаскивание некоторых из вновь введенных концептов в другие ветви дерева признаков).

5. Результаты и их обсуждение

Текущее состояние проекта таково. Выполнен анализ словарных определений — в соответствии с описанием в разделе 3 — трех словарей: Российский энциклопедический словарь (свыше 26 тыс. определений), русская Википедия (свыше 42 тыс.; по возможности, исключены персоналии), толковый словарь Т. Ф. Ефремовой (свыше 120 тыс.). Систематически организованных количественных оценок точности определения родового термина пока не делалось. Качественная оценка, полученная путем экспертного оценивания самими разработчиками результатов обработки случайно выбранной 1000 определений в первом из названных словарей дала — для «доброкачественных» определений — около 95 % правильных решений, для всех определений — порядка 85 %. Реализована и находится в процессе содержательной отладки технология энциклопедического импорта, описанная в разделе 4. Опыт, полученный в ходе отладки показал, что уже в своем настоящем виде предлагаемая технология существенно упрощает и ускоряет как отбор терминов для онтологии, так и собственно процесс ввода.

Перспективы дальнейшего развития технологии: использование более точных средств синтаксического анализа, расширяющих возможности точной идентификации онтологических характеристик терминов, в частности, обработка в определенных ситуациях межсегментных связей; формализация терминов других категорий: *наименования признаков и процессы*; поиск и формализация в тексте определения по возможности, всех онтологически значимых характеристик термина; использование наряду с *rule-based* процедурами статистических метрик.

Литература

1. *Buitelaar P., Cimiano P.* Ontology learning and population: bridging the gap between text and knowledge. // Series: Frontiers in artificial intelligence and applications, v. 167. — Amsterdam ; Washington, DC : IOS Press, 2008.
2. *Chodorow M. S., Byrd R. J., Heidorn G. E.* Extracting semantic hierarchies from a large on-line dictionary. Proceedings of the 23rd Annual Conference of the Association for Computational Linguistics, Chicago, 1985, pp. 299–304.
3. *Ide N., Véronis J.* Extracting knowledge bases from machine-readable dictionaries : Have we wasted our time? // KB&KS'93 Workshop,Tokyo. — 1993. — pp. 257–266
4. *Ponzetto S. P., Strube M.* Deriving a large scale taxonomy from Wikipedia // Proceedings of the 22nd Conference on the Advancement of Artificial Intelligence, Vancouver, B. C., Canada, 22–26 July 2007, pp. 1440–1445.
5. *Navigli R., Velardi P.* From Glossaries to Ontologies: Extracting Semantic Structure from Textual Definitions (в [1], pp. 71–87)
6. *Ide N., Veronis J.* Refining Taxonomies Extracted from Machine-Readable Dictionaries. // Hockey, S.,Ide, N. (eds.) *Research in Humanities Computing II*, Oxford University Press, 2003
7. *DODDLE Project*. A Domain Ontology rapiD DeveLopment Environment. URL: http://doddle-owl.sourceforge.net/en/
8. *Рубашкин В. Ш., Капустин В. А.* Использование определений терминов в энциклопедических словарях для автоматизированного пополнения онтологий // XI Всероссийская объединенная конференция «Интернет и современное общество» — СПб., 2008.
9. *Российский* энциклопедический словарь // М.: Большая Российская энциклопедия, 2001
10. *Рубашкин В. Ш.* Онтологии — проблемы и решения. Точка зрения разработчика // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2007». — М.: Издательский центр РГГУ, 2007
11. *Рубашкин В. Ш.* Представление и анализ смысла в интеллектуальных информационных системах. // М.: Наука, 1989
12. *Рубашкин В. Ш., Пивоварова Л. М.* Онторедактор как комплексный инструмент онтологической инженерии // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2008». — М.: Наука, 2008
13. *Бочаров В. В., Пивоварова Л. М., Рубашкин В. Ш.* Логико-лингвистический анализ текстов определений в энциклопедических и толковых словарях. // Мегалинг-2009
14. *Система* автоматической обработки текстов АОТ. http://www.aot.ru/

Опыт корпусного исследования морфологической вариативности: варианты родительного падежа множественного числа существительных мужского рода[1]

Corpus-based study of morphological variability: variants of genitive plural masculine in Russian

Савчук С. О. (savsvetlana@mail.ru)

Институт русского языка им. В. В. Виноградова РАН

Излагаются результаты корпусного исследования одного из неустойчивых участков морфологической системы русского языка. Данные, полученные на материале Национального корпуса русского языка, сравниваются с результатами социолингвистического исследования, проводившегося в 1960-е годы, отмечаются изменения в тенденциях развития вариантности рассматриваемой группы существительных.

1. Введение

В настоящее время одним из актуальных направлений корпусной лингвистики стало применение корпусов текстов в грамматических исследованиях. Корпусные грамматики создаются для английского[2], немецкого[3], чешского языка[4], начаты эти работы и в России на базе Национального корпуса[5]. Грамматическое описание, ориентированное не на ограниченные искусственные примеры, а на реальное употребление, вынуждает подвергнуть пересмотру устоявшиеся, ставшие привычными представления. Особенно ценным корпусной подход оказывается при изучении нестандартных, вариативных явлений, которые традиционно относятся к периферии грамматики и составляют область грамматической стилистики и культуры речи. Обширный корпусной материал дает возможность не только зафиксировать наличие вариантов, но и оценить их соотношение в статике и динамике, установить их функциональное распределение и зависимость от социолингвистических факторов. Все это способно уменьшить субъективность суждения о характере вариантов и в конечном счете сделать более убедительными и обоснованными нормативные рекомендации о выборе того или иного варианта.

Круг морфологических вариантов достаточно хорошо известен и остается постоянным на протяжении длительного периода. Однако здесь происходят постоянные микроскопические изменения, не заметные невооруженным глазом, которые меняют границы этого круга и вызывают разноречивые рекомендации в нормативной литературе. В настоящей работе приводятся результаты корпусного исследования одной из активных точек вариативности в морфологической системе русского языка — формы родительного падежа множественного числа существительных мужского рода[6]. Форма родительного падежа множественного числа занимает

[1] Работа выполнена при поддержке: Программы ОИФН РАН «Текст во взаимодействии с социокультурной средой: уровни историко-литературной и лингвистической интерпретации»; проекта «Русский язык XVIII в.: корпусные исследования лексической и морфологической вариативности и словаря» в рамках Программы фундаментальных исследований Президиума РАН «Историко-культурное наследие и духовные ценности России»; РФФИ (грант 08-06-00371-а).

[2] Francis et a. 1996; Francis et al 1998; Biber et al 1999; Huddleston & Pullum 2002.

[3] Zifonun et al. 1997; Grammis http://hypermedia.ids-mannheim.de/

[4] См.: http://mam.ujc.cas.cz/bibliografie.php

[5] Корпусные исследования по русской грамматике 2009.

[6] Перспективы использования НКРЯ в изучении грамматических норм обсуждались в работах Гришина, Савчук 2007; Савчук, Гришина 2008; Савчук 2009.

второе место (после именительного множественного) по частоте употребления вариантных падежных форм существительных[7]и объединяет более 200[8], по другим данным, более 300[9] существительных определенных семантических разрядов.

Источником вариативности в этой точке парадигмы послужила перестройка системы склонений, которую считают одним из ранних и последовательно проведенных процессов в истории русского именного склонения (Марков 1992, 84), а также смешение типов склонения во множественном числе. Начиная с XII в., существительные мужского рода бывшего склонения на *о* (*вълкъ*) объединялись в один словоизменительный класс с существительными мужского рода бывшего склонения на *й* (*домъ*). В результате конкуренции флексий одного и того же падежа побеждала какая-то одна, более «перспективная». Для род. мн. более перспективной оказалась флексия *-овъ*, оформлявшая бывшее склонение на *й*, несмотря на то, что оно представлено гораздо меньшим количеством слов. Почему?

Флексия *-ов* формально противопоставляет словоформы род. мн. и им. ед., омонимичные в склонении на *о* (*злой волкъ — стая волкъ*), и следовательно, лучше выполняет смыслоразличительную функцию. Процесс замены у существительных бывших о-основ исконной нулевой флексии флексией *-ов* в литературном языке и в диалектах, с одной стороны, и в разных группах существительных, с другой, протекал по-разному. Не закончился он и в настоящее время. В результате в современном русском языке имеется три окончания для формы родительного множественного: *-ов*, *-ей* и ∅. Закономерности их распределения были сформулированы Р. Якобсоном (Jakobson, 1956) и приняты во многих грамматических описаниях (РГ 1980, 498; Еськова, 1994, 386; Andrews 2001, 34): если в форме именительного падежа единственного числа окончание нулевое, то во множественном числе не нулевое, а *-ов* или *–ей*, и наоборот, ненулевым окончаниям исходной формы соответствуют нулевое окончание в род. мн. Таким образом, для большинства существительных соотношение форм им. ед. и род. мн. выглядит следующим образом: *стен-а — стен-∅*, *вин-о — вин-∅, стон-∅ — стон-ов, конь-∅ — кон-ей, ступень-∅ — ступен-ей*. Однако из этой общей схемы есть исключения двоякого рода: 1) нулевому окончанию в им. ед. соответствует нулевое в род. мн. и 2) ненулевому в исходной форме — ненулевое в род. мн., как единственно возможное в литературном языке или как один из вариантов.

Им. ед.	Ожидаемая форма	Реальная форма
мор-е	*морь	мор-ей
колен-о	колен-∅	колен-∅ и колен-ей
мест-о	мест-ов	мест-∅ и место-ов (прост.)
каланч-а	*каланч	каланч-ей
яблон-я	яблонь-∅	яблонь-∅ и яблон-ей (нелит.)
глаз-∅	*глаз-ов	глаз-∅
солдат-∅	солдат-ов	солдат-∅ и солдат-ов (прост.)
гусар-∅	гусар-ов	гусар-∅ и гусар-ов

Помимо этих вариаций существуют вариации, связанные с ограничением образования род. п. мн. ч. (*кочерга — кочерег*), формы, образуемые от видоизмененной основы (*опенок — опят* и *опенков*), особенности образования форм род. п. существительных Pluralia tantum. Все это объясняет популярность данной формы, которой она пользуется у исследователей грамматики, как отечественных, так и зарубежных. Представителей генеративного направления эта форма привлекает своей непредсказуемостью, и побуждает совершенствовать модели описания (Bailyn&Nevins 2008, Pertsova 2005). В отечественной лингвистике вариативность формы род. мн. рассматривалась как в синхронических описаниях (Горбачевич 1978; Граудина 1964; Зализняк 1967; РГ 1980), так и в историческом (Булаховский 1953, 1954; Марков 1992, Обнорский 1931), и в нормативно-стилистическом аспекте (см. словари и справочники ГПРР, Грамм 1977/2003, Еськова 1994, Розенталь, СТ, ТС, Чернышев 1911 др.). Кроме того, этот участок колебаний нормы в числе других стал предметом массового обследования состояния норм методом анкетирования носителей русского языка, проводившегося в 1964–1965 гг. XX в. Результаты, полученные на основе обработки около 4000 анкет, нашли отражение в четырехтомной коллективной монографии «Русский язык и советское общество» (М., 1968), книге «Русский язык по данным массового обследования» (М., 1974), в работах Воронцова 1978 и др. и в дальнейшем учитывались при разработке нормативных рекомендаций.

Таким образом, к настоящему времени мы имеем возможность во-первых, используя корпусной материал, сравнить состояние вариантов в этой точке нормы на нескольких временных срезах; во-вторых, рассмотреть изменения, накопившиеся за последние 40 лет; в-третьих, сопоставить данные, полученные корпусными методами, с данными анкетирования, описанными в работах 1960–1970-х годов, оценить степень реализации сделанных прогнозов; в-четвертых, сравнить эмпирические данные с рекомендациями словарей и справочников. В рамках данной статьи мы ограничимся рассмотрением вариативности форм род. п. существительных мужского рода.

[7] Граудина 1980, 183.

[8] Граудина 1964.

[9] ТС; Горбачевич 1978.

В литературном языке вариативность окончаний род. п. мн. ч. этой категории существительных выражается в следующем.

1. Формы с окончанием *-ов* (*-ев*) являются стандартными и объединяют большинство существительных мужского рода с основой на твердый согласный, *j*, шипящий или *ц*: *домов, столов, волков, воробьёв, месяцев.*
2. Формы с окончанием *–ей* являются стандартными для существительных с основой на мягкий согласный: *коней, журавлей, кораблей*.
3. Формы с нулевым окончанием являются нестандартными и характерны для существительных с основой на твердый согласный, которые относятся к нескольким семантическим группам.
   3.1. названия лиц по принадлежности к какой-либо совокупности: этнической (*англичане — англичан, армяне — армян, буряты — бурят)*; религиозной (*христиане — христиан, мусульмане — мусульман)*; воинским соединениям (*гусары — гусар, драгуны — драгун, уланы — улан)*;
   3.2. названия единиц измерения в сочетании с количественными словами: (несколько) *аршин, грамм, ватт, ампер, мегабайт, килобит, рентген, гектар*. Правда, в этом случае говорят об особой счетной форме, а не о родительном падеже.
   3.3. названия некоторых парных предметов: *чулок, ботинок, сапог, брюк, погон, рожек.*
   3.4. названия овощей и фруктов (здесь форма с нулевым окончанием признается нормой только в качестве варианта в разговорной речи): (килограмм) *баклажан, гранат, помидор, апельсин, мандарин.*

Большинство существительных группы 3 образует нулевую форму вариативно. Литературной нормой может признаваться один из вариантов, а другой отвергаться: *сапог — сапогов* (прост.), *носков — носок* (не рек.). Либо оба варианта могут признаваться, но при этом получать разную стилистическую оценку: *апельсинов — апельсин* (разг.). Нормативные оценки конкретных вариантов подвержены влиянию общественного вкуса, что не может не отражаться на скорости процессов их изменений. В грамматических описаниях эта группа часто характеризуется как один из очагов развития аналитизма в русской морфологической системе (РЯСО, 66) или как свидетельство действия закона речевой экономии (Валгина 2001, 175; Розенталь 1965). На последнее утверждение можно возразить следующее: если экономия выражается в выборе короткого окончания, и «*килограмм мандарин*» экономнее «*килограмма мандарин**ов***», то почему подобной экономии не подвергается «*килограмм мандаринчик**ов***»? Что касается нарастания аналитических тенденций в морфологии, то применительно к данной группе существительных оно должно проявляться в увеличении доли нулевых форм в паре конкурирующих вариантов и в расширении круга существительных с предпочтительной ∅ формой. В лингвистической литературе можно найти разные точки зрения на динамику соотношения вариантных флексий род. мн. Согласно одной из них, поддерживаемой историками языка, экспансия форм на *-ов*, начавшись в XII в., продолжается и в настоящее время (Марков 1992, Обнорский 1931). Сторонники другой точки зрения предсказывают обратное направление развития, отмечая укрепление форм с нулевой флексией (РЯСО; Активные процессы 2008). Корпусное исследование вариантных форм род. мн. отдельных групп существительных может пролить свет на эту проблему.

2. Помидор и помидоров: вариантные формы названий овощей и фруктов

> Я: Дим, как правильно сказать, килограмм *апельсин* или *апельсинов*?
> Дима (подумав): Килограмм цитрусовых.
>
> Домашний урок грамматики (из блогов)

Границы этой группы не вполне четки, поскольку количественный состав рассматриваемых слов колеблется в разных словарях и справочниках. Отличаются и рекомендации относительно вариантов. В последней трети XX в. эта группа представлялась ареной наиболее активной борьбы флексий *-ов* и ∅ (Воронцова 1976, 135). Сведения по основным нормативным изданиям приведены в таблице 1.

Как видно из таблицы, рекомендации в отношении вариантов в разных изданиях расходятся: для лексемы *баклажан* нулевой вариант род. мн. допускается во всех стилях речи, а согласно ГПРР, даже признается основным. Для слов *апельсин*, *мандарин*, *абрикос*, *помидор*, *томат* нулевой вариант допускается только в устной (или разговорной) речи (ТС, Розенталь, Скворцов), а Еськовой отвергается как неправильный (*апельсин, абрикос, ананас)* или не рекомендуемый (*помидор, томат*). Помимо приведенных в таблице слов, в словарь ГПРР включены *гранат* (с нулевым вариантом как единственно правильным!), *патиссон* (с -ов и ∅), *лимон* (-ов), *артишок* (-ов), *шампиньон* (-ов и ∅), а в Грамм — *нектарин* (-ов).

Таблица 1

Лексема	Грамм	Еськова	ГПРР	ТС	СТ	Роз	Валг	Сквор
апельсинов/ апельсин	-ов	-ов, !неправ ∅	-ов и ∅(разг)	-ов и доп. ∅	-ов и доп. (в УР) ∅	-ов, в УР ∅	-ов и ∅	-ов, в РР ∅
баклажанов/ баклажан	-ов [∅]	-ов и доп ∅	∅	-ов и доп. ∅	-ов и доп. (в УР) ∅		∅ и -ов	-ов и доп. ∅
помидоров/ помидор	-ов [∅]	-ов !не рек ∅	-ов и ∅	-ов и разг. ∅	-ов и доп. (в УР) ∅	-ов, в УР -∅	∅ и -ов	-ов, в РР ∅
мандаринов/ мандарин	-ов	-ов !не рек ∅	-ов и ∅	-ов и разг. ∅	-ов и доп. (в УР) ∅	-ов, в УР -∅	∅ и -ов	-ов, в РР ∅
абрикосов/ абрикос	-ов	-ов, !неправ ∅	-ов	-ов и разг. ∅	-ов не рек. ∅			-ов в РР доп. ∅
ананасов/ ананас	-ов	-ов !неправ ∅	-ов	-ов не ∅	-ов не рек ∅		-ов	
томатов/ томат	-ов	-ов !не рек ∅	-ов	-ов	-ов не рек. ∅	-ов, в УР ∅		
бананов/ банан	-ов	-ов	-ов	-ов не ∅	-ов не рек. ∅		-ов	-ов, не ∅

В процессе корпусного исследования для каждого из существительных было проанализировано употребление ∅ и *-ов*-варианта род. мн. в четырех подкорпусах: I. XIX — 1-я пол. XX в. (68 млн словоупотреблений); II. — 2-я пол. XX в. (97,4 млн с/у); III. — корпус СМИ 2000 годов (113 млн с/у); IV. — устные тексты (8,5 млн с/у). Было отобрано 11 существительных, обнаруживших наличие вариантов хотя бы в одном из подкорпусов. Относительные частоты встречаемости вариантов на 1 млн словоупотреблений приведены в Таблице 2.

Таблица 2

Словоформа	XVIII–XX-1	XX-2	СМИ 2000	Устный
баклажанов	0,100	0,790	0,805	0,120
баклажан	0,060	0,050	0,027	0,120
помидоров	0,560	2,870	3,480	1,200
помидор	0,040	0,090	0,020	0,600
апельсинов	2,120	1,070	1,270	0,350
апельсин	0,040	0,000	0,000	0,240
мандаринов	0,210	0,400	0,500	0,120
мандарин	0,030	0,000	0,010	0,120
абрикосов	0,210	0,650	0,410	0,120
абрикос	0,000	0,000	0,020	0,120
лимонов	0,680	0,510	0,470	0,120
лимон	0,000	0,000	0,000	0,120
бананов	0,490	1,440	1,350	2,100
банан	0,000	0,000	0,000	0,000
томатов	0,070	1,080	1,500	0,000
томат	0,000	0,000	0,000	0,000
ананасов	0,470	0,390	0,380	0,820
ананас	0,000	0,000	0,000	0,000
шампиньонов	0,260	0,470	1,060	0,000
шампиньон	0,000	0,000	0,000	0,000
гранатов	0,030	0,040	0,040	0,000
гранат	0,030	0,000	0,000	0,120

Из таблицы видно, что частотность нулевого варианта (представленного в текстах XVIII–XIX в.), в текстах 2-й пол. XX в. уменьшается по сравнению с текстами предшествующего периода, а у слов *апельсин*, *мандарин*, *гранат* падает до нуля. Исключение составляет вариант род. мн. *помидор*, для которого в корпусе 2-й пол. XX в. отмечен рост частоты[10], хотя в корпусе СМИ она значительно ниже[11]. Что касается данных корпуса устной речи, то частотность нулевой формы в ней выше, чем в письменных текстах, однако ни в одном слове нет того резкого преобладания этого варианта над вариантом *-ов*, которое отмечалось в исследованиях 1960-х годов[12].

Проследим, как изменяется характер количественных оппозиций внутри вариантных пар. Корпусные данные были сопоставлены с результатами, полученными при анкетировании в 1964–1965 гг., и записями в магазинах 1962–1963 гг. Кроме того, для анализа употребления в современной разговорной речи были привлечены данные блогов, поскольку данные устного корпуса, в силу своей малочисленности, кажутся не вполне представительными. В блогах были проанализированы употребления

[10] Большая часть контекстов относится к бытовой сфере (письма, дневники) и художественной литературе, в которой художественная речь имитирует разговорную.

[11] Эти данные не соответствуют выводам о том, что к 1960-м годам употребительность нулевого окончания род. мн. возросла (Граудина 1964).

[12] «Форма с нулевой флексией настойчиво утверждается у названий овощей и фруктов: баклажан, помидор, мандарин, апельсин. По записям в магазинах (1962–1963 гг.) отмечены следующие колебания форм: абрикос — абрикосов 45/5, апельсин — апельсинов 100/0, банан — бананов 11/39, баклажан — баклажанов 100/0, гранат — гранатов 48/2, мандарин — мандаринов 47/3, помидор — помидоров 396/6. Судя по выборке, в устной речи преобладает форма с нулевой флексией» (РЯСО, 82).

Таблица 3

Доля *-ов*	Магазины 1962–1963	Анкета 1964–1965*	Блоги 2010	НКРЯ XVIII–XX-1	НКРЯ XX-2	НКРЯ СМИ XXI
баклажанов	0,000	0,365	0,570	0,636	0,940	0,970
помидоров	0,015	0,405	0,620	0,927	0,959	0,985
мандаринов	0,060	0,483	0,670	0,875	1,000	0,982
апельсинов	0,000	0,618	0,860	0,980	1,000	1,000
абрикосов	0,100		0,830	0,930	1,000	0,958
гранатов	0,040		0,640	0,500	1,000	1,000
бананов	0,780		0,970	1,000	1,000	1,000

* Данные приводятся по РЯСО, 82–83 и Воронцова, 140–142.

вариантов в контекстах с количественным и качественным значением — со словами «килограмм (кило) Х/Хов» и «из Х/Хов»[13]. Результаты представлены в таблице 3.

Цифры обозначают долю варианта *-ов* в общем количестве вариантов род. мн. Например, в корпусе СМИ 2000-х годов форма *баклажанов* встретилась 91 раз, а форма *баклажан* — 3 раза, следовательно, доля варианта *-ов* к общему количеству употреблений род. мн. (94) составляет 0,97, или 97 %, а доля варианта ∅ — соответственно 0,03, или 3 %. Поскольку эти величины связаны между собой, достаточно провести вычисления для одного из вариантов. Единица (или 100 %) означает, что нулевой вариант соответствующей лексемы отсутствует, пустые клетки означают отсутствие данных. Динамика соотношения вариантов показана на Рис.1 и Рис. 2.

График на Рис. 1 демонстрирует тенденцию к явному и убедительному предпочтению варианта *-ов* в форме род. мн. В письменной речи он практически вытесняет нулевой вариант, а в разговорной речи, в сопоставлении с данными 1960-х годов, также обнаруживает тенденцию к росту, вопреки делавшимся 40 лет назад прогнозам. Эта тенденция характеризует все существительные данной группы, хотя слабая привязанность к отдельным лексемам, отмеченная в (РЯСО, 82) сохраняется: у слов *баклажан* и *помидор*, а в разговорной сфере еще и у слов *мандарин* и *гранат* доля флексий *-ов* чуть ниже, а доля нулевых флексий, соответственно, выше, чем в остальной группе.

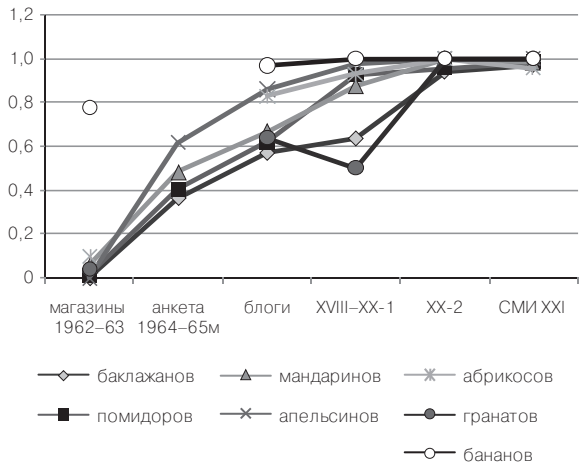


Рис. 1

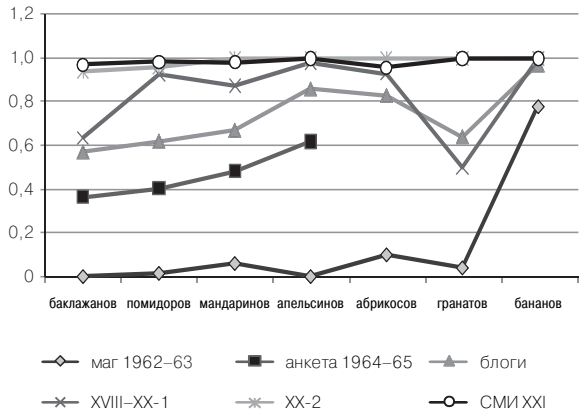


Рис. 2

На рис. 2 видно, что в каждом подкорпусе рассматриваемые существительные образуют однородную группу, проявляющую сходные грамматические свойства[14].

[13] Контекст «кило Х/Хов» может показаться наиболее приближенным к непринужденному устному общению, однако на деле в количественном отношении и по соотношению вариантов он мало отличался от контекста «килограмм Х/Хов», поэтому в расчетах данные, полученные для обоих употреблений, суммировались. В отличие от «килограмма», контекст «тонна Х/Х-ов», гораздо менее представленный в блогах, встречается в контекстах публицистического содержания и характеризуется иным соотношением вариантов. Разумеется, материал блогов можно использовать лишь для установления соотношений и выявления тенденций, поскольку количественные данные, полученные простым поиском по блогам, имеют большую погрешность по сравнению с корпусными данными (повторяющиеся тексты, принципиально открытый и изменяющийся состав текстов, что, кстати, затрудняет или делает невозможной перепроверку результатов). Уменьшение погрешности требует дополнительной ручной или программной обработки результатов.

[14] Низкая доля варианта гранат в корпусах текстов XVIII-XIX вв. и блогах есть следствие низкой частотности

Корпусной материал не подтвердил высказывавшиеся суждения о том, что нулевые формы предпочитают количественные контексты (сочетания с числительными и количественными словами): они обнаруживают такое же безразличие к контексту, как и формы на *-ов*[15]. Ср.: *Посадили штук 50 огурцов в теплицу, 50 помидор в открытый грунт, 5 перцев, кабачки, зелень, горох.* [Письмо семейное (2003)]. — *Это у вас там она садится, а мы ей тут садиться не даем, у нас, в стране вечнозеленых помидор и непуганых браконьеров…* [Виктор Астафьев. Царь-рыба (1974)]. *А ты тоже три дня ел суп из перловки и зелёных помидор?* [Василий Гроссман. Жизнь и судьба, часть 2 (1960)]. Однако обследование материала блогов дает основание для предположения о том, что в разговорной речи нулевые формы все же тяготеют к количественным контекстам[16].

Таким образом, если на основании массового обследования, проведенного в 1960-е годы, были сделаны выводы о «медленном, но неуклонном процессе смены традиционной нормы», который выражался в том, что в момент акта речи предпочитается форма с нулевой флексией, а не форма на *-ов* (РЯСО, 83), то в начале XXI века ситуация представляется несколько иной[17]. Медленные процессы в изучаемой группе существительных (которые относятся к конкретной лексике, активно используемой прежде всего в бытовой сфере, в художественной литературе и публицистике, связанной с бытовой тематикой) идут в направлении поляризации форм род. мн.: формы на *-ов* занимают доминирующее положение в сфере книжно-письменной речи, вытесняя нулевые формы в сферу разговорной речи. В сфере неформального общения, устного и письменного (существующего прежде всего в форме электронной коммуникации), флексии на *-ов* также преобладают и осознаются как нейтральные и более правильные, однако доля нулевых форм там значительно выше, чем в книжно-письменных сферах и сопоставима с данными опросника 1964–65 годов. В этом процессе можно усмотреть проявление такого фактора, как продолжающаяся функционально-стилистическая дифференциация литературного языка, разграничивающая и противопоставляющая книжно-письменные сферы с их системой кодифицированных норм и разговорную сферу, в которой также формируется своя норма неформального общения. Таким образом, описательная грамматика должна объективно фиксировать такое положение дел.

Что касается нормативных рекомендаций (см. таблицу 1), то материал корпуса не дает оснований для дифференцированных рекомендаций большинства пособий в отношении разных лексем, отделяющих *баклажаны* от *помидоров*, *апельсинов* и *абрикосов*. Наибольшее соответствуют результатам корпусного исследования рекомендации ТС и Грамм, в которых нормативно закрепленной и рекомендуемой называется форма на *-ов*, а разговорная речь (и авторские задачи в художественной речи и публицистике) оставляют возможность выбора между нормативным и разговорным вариантом.

3. Гардемарин и гардемаринов: Вариантные формы названий военных

> «Если бы в 1910 году синий кирасир услышал бы форму *кирасиров*, то он отсек бы обидчику голову!» (офицер в отставке).
>
> Из ответов на вопрос морфологического вопросника[18].

Группа названий лиц по принадлежности к воинским соединениям представляет собой закрытый список, включающий 14 существительных[19]. Варианты род. мн. всех существительных были рассмо-

этого слова в целом (основная часть вхождений формы гранатов относится к лексеме гранат «драгоценный камень», а формы род. мн. гранат — к лексеме граната «оружие»). В письменных текстах второй половины XX века вариант гранат можно считать окончательно вышедшим из употребления. Напротив, слово банан, для которого (единственного из всей группы названий плодов) в магазинных записях 1960-х годов было зафиксировано преобладание формы род. п. на -ов, в начале XXI века вдруг обнаружило некоторую активность (до 3 %) нулевой формы (в материале блогов встретилось 14 употреблений формы: «варенье из банан», «шашлык из банан» и под.).

[15] См. об этом в связи с анализом группы существительных-этнонимов в (Timberlake 2004, 138).

[16] Так, в сочетании с предлогом «из помидор — из помидоров» доля нулевой формы составляет 28 %, а в контексте «килограмм (кило) помидор — килограмм (кило) помидоров» ее доля составляет 38 %; «из баклажан» — 33 %, а «килограмм (кило) баклажан» — 43 %. Для других лексем соотношение следующее: апельсин 7 % и 14 %, мандарин — 13 % и 33 %, банан — 1,5 % и 3 %, шампиньон — 0,1 % и 3 %, однако форма абрикос чаще встретилась в контексте «из абрикос» — 29,4 % против 17 %. Интересно, что именно лексема абрикос обнаружила рост вариантов с нулевой флексией (не только в блогах, но и письменной речи) вопреки утверждению о том, что наиболее расположены к нулевой флексии имена с основой на сонорные (Горбачевич, 1978, 191).

[17] Изменению картины способствуют как сама изменившаяся речевая практика, так и новые методы и инструменты исследования, дающие возможность обрабатывать огромные массивы текстов различных речевых сфер.

[18] Русский язык и советское общество: Морфология и синтаксис современного русского литературного языка. / Под ред. М. В. Панова. М.: Наука, 1968. С. 86.

[19] В ГПРР к этой группе отнесено существительное канонир, однако в корпусном материале не выявлено ни одного случая вариантной формы род. мн. этого слова, и потому оно не рассматривается.

трены в письменных текстах разных периодов: XVIII в., XIX в., 1-й пол. XX в., 2-й пол. XX в.[20]В таблице 4 приводятся результаты обследования — частота встречаемости словоформ на 1 млн словоупотреблений текста.

Таблица 4

Словоформа	XVIII	XIX	XX-1	XX-2
солдат	28,80	>38,50	>62,50	>41,10
солдатов	0,38	0,19	0,28	0,06
партизан	0,00	0,27	>7,50	>3,60
партизанов	0,00	0,88	0,13	0,04
рекрут	8,90	3,70	0,13	0,01
рекрутов	0,39	1,00	0,50	0,27
кадет[1] (военный)	0,39	1,5	1,35	0,10
кадетов[1] (военный)	3,10	0,35	0,33	0,26
кадет[2] (член партии КД)	0,00	0,00	0,45	0,03
кадетов[2] (член партии КД)	0,00	0,00	4,80	0,49
гренадер	3,80	1,42	0,85	0,09
гренадеров	0,39	0,85	0,83	0,13
гардемарин	0,00	0,39	0,33	0,02
гардемаринов	0,00	0,58	0,25	0,12
гусар	4,20	5,00	1,25	0,23
гусаров	0,39	1,46	0,38	0,10
карабинер	1,90	0,15	0,00	0,00
карабинеров	0,77	0,23	0,18	0,10
драгун	2,30	3,03	0,95	0,09
драгунов	0,00	0,69	0,08	0,01
кирасир	0,00	0,77	0,75	0,09
кирасиров	0,00	0,54	0,08	0,07
улан	0,00	1,65	0,65	0,07
уланов	0,00	1,03	0,13	0,09
янычар	4,20	0,88	0,18	0,14
янычаров	0,39	0,08	0,13	0,00
рейтар	0,00	0,15	0,00	0,07
рейтаров	0,00	0,04	0,05	0,08

Все слова изучаемой группы, за исключением слов *солдат* и *партизан*, имеют сравнительно невысокую частотность в современных текстах (не больше 0,5 словоупотреблений на миллион) и, по большей части, относятся к устаревшей лексике. В зависимости от соотношения вариантов окончания род. мн. всю группу можно разбить на 3 подгруппы.

В подгруппу 1 входят 2 слова, относящиеся к активному лексикону: *солдат* и *партизан*. Каждое из них имеет единственный кодифицированный вариант с нулевым окончанием, а варианты с окончанием *-ов* появляются в текстах спорадически или используются в художественных текстах для имитации народной речи или просторечия. Например, *Им как-то не с руки отказываться от бесплатной рабочей силы в лице солдатов-призывников.* [Анжелика Матвеева. Не отступать и не сдаваться! // «Красноярский рабочий», 2003]. Конкуренция вариантов род. п. слова *солдат* в целом завершилась в XVIII в., а слова *партизан* — к началу XX в.[21]

В подгруппу 2 входят слова, которые, снизив свою частотность в современном языке, все же не вышли из употребления или по тем или иным лингвистическим или экстралингвистическим причинам повысили активность в конце XX в.: *кадет*[1] ('воспитанник военного учебного заведения' в связи с возрождением таких заведений), *кадет*[2] ('член политической партии конституционных демократов' в связи с интересом к событиям начала XX в.), *гардемарин*, *гусар* (герои кинематографа), *карабинер* ('солдат итальянской армии'), *рекрут*, *гренадер* (расширили значения).

Подгруппу 3 составляют слова-историзмы, которые обозначают реалии, вышедшие из употребления, — называют подразделения царской армии. Их можно встретить в текстах исторической тематики: *драгун*, *кирасир*, *улан*, *рейтар*, *янычар*.

Соотношение вариантов с нулевым и ненулевым окончанием род. мн. в рассматриваемой группе коррелирует с распределением по выделенным подгруппам. В подгруппах 2 и 3 на протяжении XVIII–XX веков продолжалась конкуренция вариантов с нулевым и ненулевым окончанием, причем

[20] Результаты см. Савчук 2009.

[21] Интересно, что это слово первоначально появилось в русском языке в XVIII в. в значении 'приверженец, сторонник какой-либо партии, группировки' (первая фиксация в корпусе относится к 1715 г.) и оформлялось в род. п. исключительно с помощью окончания -ов. Корпус дает примеры такого употребления, которое в современном языке считается устаревшим: Лопухов видел вещи в тех самых чертах, в каких представляются они всей массе рода человеческого, кроме партизанов прекрасных идей. [Н. Г. Чернышевский. Что делать? (1863)]. Он видел, что при этом общем подъеме общества выскочили вперед и кричали громче других все неудавшиеся и обиженные; главнокомандующие без армий, министры без министерств, журналисты без журналов, начальники партий без партизанов. [Л. Н. Толстой. Анна Каренина (1878)]. В период войны 1812 года активизировалось новое значение 'начальник легкого, летучего отряда, вредящего внезапными покушениями с тылу, с боков' (Даль), первоначально с формой род. мн. на -ов. Ср.: Господствующая мысль партизанов той эпохи долженствовала состоять в том, чтобы теснить, беспокоить, томить, вырывать, что по силам, и, так сказать, жечь малым огнем неприятеля без угомона и неотступно. [Д. В. Давыдов. Дневник партизанских действий 1812 года (1830–1835)]. Это последнее значение укрепилось в языке (с положительной коннотацией 'участник народной войны с захватчиками') и приобрело вариантную нулевую форму род. мн. Первая фиксация нулевой формы в корпусе относится к середине XIX в.: Граф Клингспор шел берегом, а полковник Фиянд с отрядом партизан действовал на правом фланге Раевского, зашел в тылу, захватил русский магазин, и взбунтовал всех жителей в тыл нашего отряда. [Ф. В. Булгарин. Воспоминания (1846–1849)].

в XX веке достаточно активно. Однако к настоящему времени результаты этой конкуренции в разных подгруппах различны. На Рис. 3 показано изменение доли вариантов с окончанием ∅ в словах подгруппы 2.

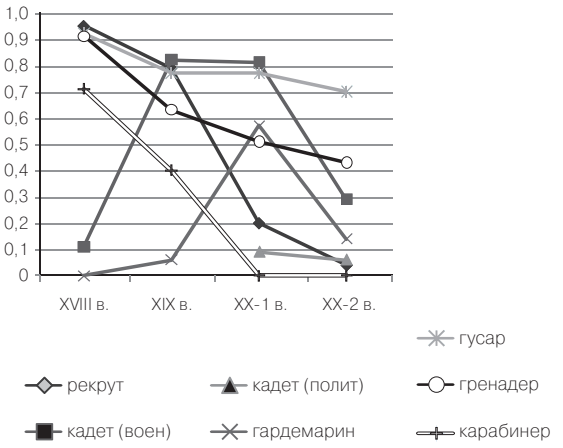


Рис. 3

Как видим, в группе слов, которые продолжают активно употребляться в новых контекстах, доля вариантов с нулевым окончанием сокращается, а доля вариантов с окончанием *-ов*, соответственно растет. У всех существительных этой группы, за исключением *кадетов*, в XVIII веке преобладало нулевое окончание, а к XX веку формы с нулевым окончанием количественно преобладают только у существительного *гусары*. Вероятно, эти формы поддерживаются устойчивыми контекстами: «Эскадрон гусар летучих» и пр. В остальных случаях, если слово активизируется в речи в прежнем значении, но в новом контексте (например, в кинематографе, как *гардемарины* в многосерийном фильме «Гардемарины, вперед!») или в новом значении (*кадеты* как ‘воспитанники военных учебных заведений в современной России’, или как ‘спортсмены до 16 лет’, *рекруты* как ‘наемники вообще’), место род. мн. занимает вариант с *-ов*.

На Рис. 4 показано изменение доли варианта с нулевым окончанием в группе 3.

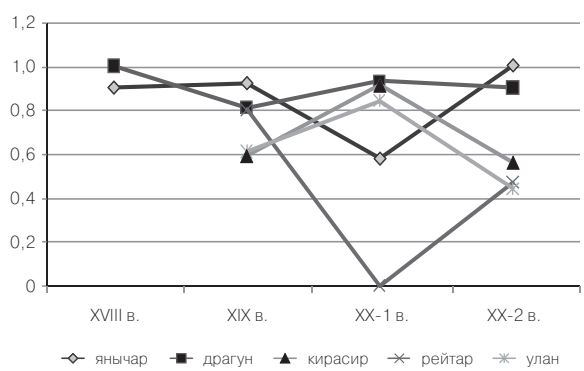


Рис. 4

У слов, относящихся к группе историзмов, вышедших из употребления в начале XX в., соотношение вариантов не развивалось в XX в. и в целом склоняется в пользу форм с нулевым окончанием. Варианты либо резко противопоставлены в количественном отношении (*янычар*, *драгун*), либо представлены в приблизительно равном соотношении (*кирасир*, *улан*). В настоящее время эти слова употребляются в текстах исторической тематики. Поскольку они относятся к пассивному запасу, их использование предполагает обращение к словарям и справочникам, что способствует сохранению традиционной формы. Интересно, что в произведении одного автора в идентичных контекстах могут присутствовать оба варианта[22].

Дополнительным подтверждением высказанного предположения послужил анализ употребления слов рассматриваемой группы в новом корпусе — СМИ 2000-х годов. Он содержит однородные тексты главным образом современной тематики, потому частотность слов пассивного словаря в них заведомо ниже, так что снижается вероятность обнаружения конкретного падежного варианта. Соотношение вариантов род. мн. существительных разных подгрупп выглядит следующим образом.

У слов подгруппы 1 (*солдат* и *партизан*) соотношение вариантов не отличается от данных, полученных по корпусу 2-ой пол. XX в. (далее — сбалансированный корпус).

Слова 2-й подгруппы представлены в корпусе СМИ в полном составе с такими показателями: *кадетов*[1] — *кадет*[1] 66/1 (ср. в сбалансированном корпусе 25/10), то есть слово даже увеличило частотность, вероятно, отчасти в связи с выходом на телеэкраны сериала «Кадеты»; *кадетов*[2] — *кадет*[2] 16/1 (в сбалансированном корпусе 48/3); в значимых количествах представлена форма род. мн. лексемы *кадет*[3] ‘спортсмен-юниор до 16 лет’: *кадетов*[3] — *кадет*[3] 16/0. Отсутствие нулевых форм весьма красноречиво. Соотношение *гренадеров* — *гренадер* 25/0 (против 13/9 в сбалансированном корпусе), при этом слово употребляется в переносном значении ‘человек высокого роста’, главным образом по отношению к спорсменам. Остальные формы представлены в следующих пропорциях: *гардемаринов* — *гардемарин* 21/0 (против 12/2 в сбалансированном корпусе); *карабинеров* — *карабинер* 29/0 (против 11/0), во всех контекстах речь идет об итальянских карабинерах; *рекрутов* — *рекрут* 30/0 (против 26/1), слово употребляется в расширенном значении не только по отношению к солдатам-новобранцам,

[22] — В обмен на четырех рейтар — одного старика. [Ю. П. Герман. Россия молодая. Часть первая (1952)] и В это время из-за угла, из-за арсенала выехал полковник Снивин в сопровождении дюжины иноземных рейтаров. [Ю. П. Герман. Россия молодая. Часть первая (1952)].

но и по отношению к работникам, спортсменам, бандитам, и даже государствам («прием новеньких государств-рекрутов» в ЕС). Формы слова *гусар* представлены в пропорции *гусаров* — *гусар* 5/25 (против 10/23 в сбалансированном корпусе), из них 13 случаев употребления в контексте «Эскадрон гусар летучих».

Из 3-й подгруппы 2 слова (*рейтар* и *улан*) вообще не представлены в корпусе СМИ в форме род. мн. У других слов группы интересующие нас формы соотносятся следующим образом: *драгун* — *драгунов* 6/0 (против 9/0), *кирасир* — *кирасиров* 1/1 (против 9/7), *янычар* — *янычаров* 7/2 (против 14/0).

Как видим, тенденция к распределению окончаний *-ов* и ∅, характерная для современных текстов, выражена в корпусе СМИ более ярко, чем в современном сбалансированном корпусе, что может объясняться большей однородностью корпуса прессы. Слова актуальные подчиняются общим правилам, закрепляют за собой стандартное окончание *-ов*, утрачивая нулевой вариант. Слова пассивного запаса вариативность сохраняют или употребляются с окончанием ∅. Не соответствуют общей тенденции слово *гусар*, у которого ∅ вариант поддерживается устойчивым контекстом, и слово *янычар*, у которого несколько увеличилось количество вариантов *-ов*.

Таким образом, в рассматриваемой группе «военных» названий в целом, как показывает материал корпуса, начиная с XVIII века наблюдается распространение флексии *-ов*, которая воспринимается как основная флексия род. п. мн. ч. у существительных мужского рода с основой на твердый согласный. Это способствует выравниванию парадигмы множественного числа, в которой формы мужского рода (с окончанием *-ов*) противопоставляются формам немужского (женского и среднего, pluralia tantum) рода (с нулевым окончанием).

Что касается нормативных рекомендаций, то разные пособия предлагают в отношении этой группы разные решения. Приведем наиболее авторитетные.

Розенталь 1965: рекомендуется нулевой вариант для всех слов.

ГПРР допускает варианты норм: предлагаются нулевые варианты для слов *гардемарины*, *гренадеры*, *рейтары*, *солдаты*, *уланы* и семантические правила распределения вариантов у остальных слов: *кадет* (‘о воспитанниках военных учебных заведений’), но *кадетов* (‘о членах конституционно-демократической партии’); *драгун*, *кирасир*, *янычар* (при собирательном значении — со словами *отряд*, *эскадрон* и пр.) и *драгунов*, *кирасиров*, *янычаров* (при обозначении отдельных лиц).

Грамм 1977/2003 и Еськова 1994: дается разрешительная помета *и* — допустимы оба варианта, причем у Еськовой первым идет нулевой вариант, в Грамм 1977/2003 — вариант с *-ов*.

Как представляется, рекомендации нулевого варианта у Розенталя и Граудиной кажутся устаревшими и не подтверждаются живой практикой словоизменения в этой группе слов, часть из которых переживает на рубеже XX и XXI вв. второе рождение. Семантические и сочетаемостные разграничения кажутся излишне тонкими и слабыми на фоне общей тенденции к грамматическому родовому размежеванию, и рекомендации, построенные на них, не соблюдались уже в XVIII и XIX веке. Ср.: … сам же взял с собою… суздальских шестьдесят *гренадеров*, сто мушкетеров, более ста стрелков, при двух пушках, и тридцать шесть воронежских *драгун*. [А. В. Суворов. Биография А. В. Суворова, им самим написанная в 1786 году]. Миних послал вперед к Яссам Кантемира с трехтысячным отрядом волохов, *драгунов* и *гусар*, а сам следовал за ним. [Н. И. Костомаров. Русская история в жизнеописаниях ее главнейших деятелей. Выпуск седьмой: XVIII столетие (1862–1875)]. Наша счастливая атака французских *кирасиров* и *драгун*. [Ф. В. Булгарин. Воспоминания (1846–1849)].

Таким образом, наиболее приемлемой и отвечающей современному состоянию нормы кажется разрешительная помета для существительных, относящихся к подгруппам 2 и 3, рекомендующая вариант с *-ов* в качестве основного и нулевой вариант в качестве дополнительного, уместного, в частности, в архаизированной речи. Именно такие рекомендации относительно рассматриваемой группы существительных предлагаются в Грамматическом словаре русского языка А. А. Зализняка.

***

Если рассматривать всю группу существительных мужского рода, испытывающих колебание в образовании форм род. мн., то несмотря на общие свойства, объединяющие их, — семантический признак количества и генетический признак иноязычного происхождения, они обнаруживают специфику взаимоотношения вариантов внутри каждой подгруппы и даже у отдельных лексем, что отмечалось в литературе (Воронцова 1978, 135–144) и что подтвердило корпусное исследование полного состава двух подгрупп. Поэтому общие выводы относительно тенденций развития вариативности во всей группе существительных мужского рода с твердой основой можно будет сделать на основании наблюдений за поведением вариантов во всех выделяемых в ней подгруппах.

Литература

1. *Булаховский 1953* — Булаховский Л. А. Курс русского литературного языка. Т. 2. Киев: Радянська школа, 1953
2. *Булаховский 1954* — Булаховский Л. А. Русский литературный язык первой половины XIX века. М.: Учпедгиз, 1954
3. *Валгина 2001* — Валгина Н. С. Активные процессы в современном русском языке. М: Логос, 2001
4. *Воронцова 1976* — Воронцова В. Л. Варианты флексии -ов и -∅ в родительном падеже множественного числа существительных мужского рода // Социально-лингвистические исследования / Ред. Л. П. Крысин и Д. Н. Шмелев. М., 1976. С. 129–144
5. *Горбачевич 1978* — Горбачевич К. С. Вариантность слова и языковая норма. Л.: Наука, 1978
6. *ГПРР* — Граудина Л. К., Ицкович В. А., Катлинская Л. П. Грамматическая правильность русской речи. М.: Наука, 1976; 3-е изд. 2004
7. *Грамм 1977* — Зализняк А. А. Грамматический словарь русского языка. Словоизменение. М.: Русский язык, 1977
8. *Грамм 2003* — Зализняк А. А. Грамматический словарь русского языка. Изд. 4-е, испр. и доп. М.: 2003
9. *Граудина 1971* — Граудина Л. К., Ицкович В. А., Катлинская Л. П. Грамматические варианты: Опыт частотного словаря. М.: Наука, 1971
10. *Граудина 1980* — Граудина Л. К. Вопросы нормализации русского языка: Грамматика и варианты. М.: Наука, 1980.
11. *Гришина, Савчук 2007* — Гришина Е. А., Савчук С. О. Национальный корпус русского языка как инструмент для изучения вариативности грамматических норм // Труды международной конференции «Корпусная лингвистика — 2008» 6–10 октября 2008 г. СПб, 2008.
12. *Еськова* — Еськова Н. А. Краткий словарь трудностей русского языка. Грамматические формы. Ударение. М.: Русский язык, 2003
13. *Еськова Н. А.* Нормы русского литературного языка XVIII–XIX. Ударение. Грамматические формы. Варианты слов. Словарь. Пояснительные статьи. М.: Рукописные памятники Древней Руси, 2008
14. *Зализняк 1967* — Зализняк А. А. Русское именное словоизменение. М.: Наука, 1967
15. *Корпусные* исследования по русской грамматике / Под ред. К. Л. Киселевой, В. А. Плунгяна, Е. В. Рахилиной, С. Г. Татевосова. М.: ПРОБЕЛ-2000, 2009.
16. *Марков 1992* — Марков В. М. Историческая грамматика русского языка. Именное склонение. Ижевск, 1992
17. *Обнорский 1931* — Обнорский С. П. Именное склонение в современном русском языке, вып. 2. Л., 1931.
18. *Орф* — Русский орфографический словарь. М., 2005
19. *Розенталь 1965* — Розенталь Д. Э. Практическая стилистика русского языка. М., 1965 // Розенталь Д. Э. Русский язык: Справочник-практикум. М.: Оникс; Мир и Образование, 2007
20. *Розенталь, Теленкова 2007* — Розенталь Д. Э., Теленкова М. А. Словарь трудностей русского языка. М.: Айрис-пресс, 2007
21. *Русская грамматика 1980* — Русская грамматика / Под ред. Н. Ю. Шведовой и др. Т. 1–2. М.: Наука, 1980
22. *РЯСО* — Русский язык и советское общество: Морфология и синтаксис современного русского литературного языка. М.: Наука, 1968
23. *Русский* язык по данным массового обследования. М.: Наука, 1974
24. *Савчук, Гришина 2008* — Савчук С. О., Гришина Е. А. Вариантность в русском языке. Проект словаря // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог» (2008). Вып.7 (14).
25. *Савчук 2009* — Savchuk S. The Russian National Corpus as a Tool for the Research on Grammatical Variability // Proceedings of the Third International Conference Grammar & Corpora Mannheim, September 22–24, 2009 (в печати)
26. *Скворцов 2007* — Скворцов Л. И. Большой толковый словарь правильной русской речи. М.; СПб: ДИЛЯ, 2007
27. *СТ* — Горбачевич К. С. Словарь трудностей современного русского языка. — СПб: Норинт, 2003
28. *Современный русский язык 2008* — Современный русский язык: Активные процессы на рубеже XX-XXI веков / Отв. ред. Л. П. Крысин. М.: Языки славянской культуры, 2008
29. *ТС* — Трудности словоупотребления и варианты норм русского литературного языка: Словарь-справочник / Под ред. К. С. Горбачевича. Л.: Наука, 1973
30. *Чернышев 1911* — Чернышев В. Правильность и чистота русской речи: Опыт русской стилистической грамматики. СПб, 1911
31. *Andrews Edna (2001)* —TheRussianReferenceGrammar. Available at: http://www.seelrc.org:8080/grammar/mainframe.jsp?nLanguageID=6
32. *Biber D., Johansson S., Leech G., Conrad S., Finegan E.* The Longman Grammar of Spoken and Written English. London: Longman, 1999.
33. *Bailyn&Nevins 2008* — Bailyn, John F. and Nevins A. Russian genitive plurals are impostors // Inflectional Identity / Ed. By A. Bachrach and A. Nevins. Oxford University press

34. *Francis et al 1996* — Francis, G.; Hunston, S. and Manning, E. Collins COBUILD Grammar Patterns 1: Verbs. London: HarperCollins, 1996.
35. *Francis et al 1998* — Francis G.; Hunston S. and Manning E. Collins COBUILD Grammar Patterns 2: Nouns and adjectives. London: HarperCollins, 1998.
36. *Huddleston 2002* — Huddleston R., Pullum G. The Cambridge Grammar of the English Language. Cambridge University Press, 2002.
37. *Jakobson R. O.* 1956/1984. «The Relationship between Genitive and Plural in the Declension of Russian Nouns». In L. R. Waugh and M. Halle (eds). Russian and Slavic Grammar: Studies 1931–1981, Berlin: Mouton Publishers. 135–140.
38. *Pertsova 2005* — Pertsova, Katya. How lexical Conservatism can lead to paradigm gaps // UCLA Working papers in Linguistics, # 11, September 2005
39. *Timberlake 2004* — Timberlake, Alan. A Reference Grammar of Russian. Cambridge, NY, 2004
40. *Zifonun et al. 1997* — Zifonun G., Hoffmann L., Strecker B. Grammatik der deutschen Sprache. 3 Bände. — Berlin/New York: de Gruyter, 1997

О возможностях автоматизации выявления связей между терминами предметной области (на примере катализа)[1]

Possibilities of automation of relationship identification between subject-domain terms (on the material of catalysis)

Саломатина Н. В. (nataly@math.nsc.ru),
Гусев В. Д. (gusev@math.nsc.ru)
Институт математики СО РАН, Новосибирск

Ильина Л. Ю. (ilud@catalysis.ru), **Кузьмин А. О.** (kuzmin@catalysis.ru),
Пармон В. Н.
Институт катализа СО РАН, Новосибирск

В рамках проблемы автоматизации построения тезаурусов предметных областей на базе текстовых подборок рассматриваются три подхода к выявлению связей между терминами: 1) построение профиля кластеризуемости наиболее значимых элементов текста, 2) формирование специфических шаблонов (образцов с переменными), 3) использование индикаторов связи.

Введение

Нижние уровни *онтологий* различных предметных областей (ПО) обычно представлены *тезаурусами*, содержащими информацию об основных *понятиях и терминах* ПО, а также *связях* между ними. Автоматизация построения онтологий *на основе текстов* ПО является актуальной задачей компьютерной лингвистики. Для формирования терминологических словарей разработано довольно много компьютерных подходов, минимизирующих затраты ручного труда [1÷4]. Гораздо меньше работ посвящено *выявлению связей* между терминами ПО [5, 6]. *Целью* данной работы является расширение спектра возможных подходов к *автоматизации* (хотя бы частичной) этого процесса.

Отличительной особенностью предлагаемых подходов является использование техники L-граммного анализа в сочетании с позиционным как на этапе формирования словаря по текстам ПО, так и на этапе выявления связей между его элементами. Термин L-грамма, по-видимому, был впервые введен Шенноном в [7] применительно к цепочке из L подряд следующих букв текста, а затем был перенесен (не совсем корректно) и на цепочки из L подряд следующих слов ($L = 1, 2, \ldots$). Именно в последнем смысле он используется в данной работе. Техника L-граммного представления разработана нами как применительно к одному, так и к группе текстов [8]. В первом случае с ее помощью выявляются всевозможные внутритекстовые повторы произвольной длины, во втором случае — межтекстовые повторы, что удобно для целей классификации. Наряду с информацией о частоте встречаемости каждой L-граммы в тексте, фиксируются места ее вхождения в текст (позиционная информация).

Привлекательными особенностями L-граммного подхода к формированию тезаурусов ПО являются: применимость к разноязычным текстам, ориентация на извлечение терминов произвольной длины, оценка их информативности путем привлечения позиционной информации, возможность формирования шаблонов для описания групп близких L-грамм и установления связей между ними. Важно отметить, что L-граммные спектры содержат не только терминологические цепочки, но и индикаторные, несущие информацию о связях между терминами.

[1] Работа выполнена при финансовой поддержке Интеграционного проекта СО РАН № 111.

1. Исходные данные и предобработка

Предметная область, используемая нами для иллюстрации предлагаемых подходов, связана с разделом химии, изучающим возможности ускорения или замедления химических реакций (катализ). Исходная подборка была представлена пятью текстами: z1: О. В. Крылов «Гетерогенный катализ» (учебник); z2: В. Б. Фенелонов «Введение в основы адсорбции и текстурологии» (учебник); z3: И. П. Мухленов «Технология катализаторов»; z4: «Лекции по катализу» (1÷15); z5: «Химическая энциклопедия» (фрагменты). Суммарный объем подборки — свыше 403 тыс. словоупотреблений. Характерные особенности подборки: наличие значительного числа синонимов, связанных с дублированием названий веществ их химическими формулами; большое число аббревиатур и сокращений, в том числе общеупотребительных слов (см. z5); вариативность в числе слов, используемых для обозначения одного и того же понятия (метанол ($L = 1$) ≡ метиловый спирт ($L = 2$)); наличие специфических терминов в каждом из источников (z1÷z5) (и только в нем), лежащих на периферии основной проблематики. Все эти факторы в значительной степени влияют на результаты и должны учитываться при автоматической обработке.

Предобработка исходных материалов состояла из следующих этапов:

- Нормализация текстовой подборки;
- Получение L-граммных характеристик [8] всей подборки для значений $L = 1, 2, \ldots, L_{max}$, где L_{max} — длина (число слов) максимальной повторяющейся цепочки в нормализованной подборке. В характеристике L-го порядка представлен полный спектр L-грамм, присутствующих в подборке, с указанием их частот встречаемости и распределения по отдельным источникам z1÷z5.
- Упорядочение L-граммных спектров при каждом значении L: а) по убыванию частоты встречаемости; б) лексикографически; в) по убыванию показателя неравномерности позиционного распределения (более детально об использовании этого показателя см. в [15]).

Указанные этапы являются общими как при формировании словаря, так и при выявлении связей между его элементами. Последующие этапы могут различаться в зависимости от преследуемой цели, но все они носят характер процедур *фильтрации* для отсеивания малоинформативной (в интересующем нас плане) части L-граммного спектра. Примерами процедур фильтрации являются: — отбор L-грамм ($L \geq 2$), удовлетворяющих *критерию устойчивости* [2]. Это основа для выделения многословных терминов и индикаторов связи;[2] — *учет частотной и позиционной информации* (отсеиваются низкочастотные L-граммы и L-граммы с равномерным позиционным распределением); — проверка наличия синтаксической связности слов в цепочке (желательна для элементов терминологического словаря, но необязательна для индикаторов связи); — учет частеречных значений (в терминологических сочетаниях преобладают существительные и прилагательные, среди индикаторов связи встречаются и глаголы) и др.

Заметим, что упомянутые выше процедуры упорядочения тоже, в некотором смысле, можно трактовать как процедуры фильтрации, позволяющие провести отсечение «избыточного» материала в нужном месте. Так, упорядочение в) призвано сдвинуть вниз L-граммы общеупотребительного толка, распределенные, как правило, равномерно по тексту. Это способствует концентрации терминоподобных L-грамм в начальной части списка. Наибольший интерес представляют L-граммы, поднявшиеся вверх в упорядочении в) по сравнению с упорядочением а). Например, в упорядочении а) слова *катализатор*, *поверхность*, *реакция*, *адсорбция* занимали соответственно, 6-е, 10-е, 15-е и 22-е место. В упорядочении в) они поднялись, соответственно, на 1-е, 5-е, 4-е и 2-е место. Таким образом, работая с упорядочением в) эксперт может существенно уменьшить объем просматриваемого материала.

2. Возможные подходы к выявлению связей между терминами

Наибольшую трудность представляет автоматизация процедуры выявления связей между элементами терминологического словаря и уточнение номенклатуры этих связей. На данный момент мы рассматриваем три возможности продвижения в этом направлении, связанные с использованием: а) профилей кластеризуемости; б) терминологических шаблонов; в) индикаторов связи.

2.1. Профили кластеризуемости наиболее значимых элементов текста

Этот аппарат ориентирован в первую очередь на выявление *ассоциативных связей* между элемен-

[2] Термином *устойчивая цепочка* мы характеризуем L-граммы ($L \geq 2$), встречающиеся в большом числе разнообразных контекстов. И, наоборот, неустойчивой считается цепочка, которая лишь единственным образом продолжается вправо или влево при всех своих вхождениях в текст. Это означает, что она не имеет самостоятельного значения и функционирует лишь в составе одной и той же (более длинной) цепочки. Детали формализации понятия *устойчивости* описаны в [8].

тами словаря. В отличие от совместной встречаемости, предполагающей позиционную близость двух (или большего количества) слов в рамках какого-либо устойчивого словосочетания, ассоциативно связанные слова (или словосочетания) могут быть разнесены друг от друга в общем случае на произвольное расстояние. Предполагается, что начальная версия словаря уже получена и представлена (в большинстве своем) *L*-граммами ($L \geq 1$), демонстрирующими неравномерное распределение в тексте. Наиболее характерное проявление неравномерности связано с кластеризацией вхождений *L*-граммы в отдельных участках текста. Статистически значимые кластеры могут быть выделены с помощью сканирующих статистик [9]. Кластеры, образуемые разными *L*-граммами, могут быть позиционно разнесены друг от друга, пересекаться друг с другом или вкладываться один в другой. Понятие *профиля кластеризуемости* было введено нами в [10], чтобы аккумулировать на одном графике информацию обо всех участках кластеризации разных *L*-грамм.

Формально, профиль кластеризуемости — это ступенчатая функция, аргументом которой является порядковый номер предложения в тексте, а значение равно числу различных кластеров, включающих в себя данное предложение. При этом в рассматриваемом предложении вовсе не обязаны присутствовать одновременно все *L*-граммы, кластеризующиеся в данном участке текста. Пики профиля кластеризуемости обычно соответствуют отдельным микротемам текста, а провалы между ними — переходу от одной микротемы к другой. Использование профилей кластеризуемости для выявления ассоциативных связей основано на предположении о том, что такого рода связи как раз и имеют место между элементами словаря, кластеризующимися в одном и том же участке текста.

Ниже на *Схеме 1* приведен фрагмент профиля кластеризуемости текста z1, охватывающий предложения с номерами от 3973 до 5287. Здесь ось абсцисс с номерами предложений направлена вниз, а ось ординат (число кластеров) — по горизонтали:

Схема 1. Профиль кластеризуемости фрагмента текста z1

слева направо. Для экономии места ось ординат представлена в нелинейном масштабе: указаны номера лишь тех предложений, на которых происходит изменение значений профиля, т. е. добавляются новые кластеры или исчезают старые. Относительно редкие случаи сохранения высоты соседних столбиков связаны с одновременным добавлением и устранением кластера (см., например, №№ 4515 и 4591).

Для наглядности профиль в каждой точке представлен набором чисел, отделенных друг от друга точкой с запятой. Количество чисел в наборе (значение профиля) соответствует числу кластеров, включающих в себя данное предложение. Сами же числа — это номера элементов словаря, по которым строился профиль. Таблица соответствий «номер–термин» представлена рядом с графиком. Опять же для упрощения картины профиль строился лишь по 80 биграммным комбинациям (упорядочение в)). Из них на рассматриваемом участке проявили себя в виде позиционных кластеров лишь 18[3].

В принципе, каждый набор чисел, соответствующий конкретной позиции текста, можно трактовать как набор *ассоциативно связанных терминов*. Нетрудно видеть, что соседние наборы сильно коррелируют друг с другом. Для позиционно разнесенных наборов корреляция падает. Например 6-ти элементные наборы, соответствующие предложениям с номерами 4787 и 5142 имеют лишь два общих термина: *каталитическая реакция* и *возбужденная молекула* (№№ 9 и 11 в таблице соответствий).

Наивысшее значение профиля на тексте z1 (длина 17859 предложений) равно 8 и зафиксировано в предложениях с номерами 2782, 3004 и 13202. Приведем для иллюстрации список терминов, кластеризующихся в районе позиции 13202: *каталитическая реакция*, *каталитическая активность*, *активный центр*, *атом металла*, *число атомов*, *частица металла*, *размер частицы*, *адсорбция* H_2. Нетрудно видеть, что ассоциативные связи между отдельными парами терминов проявляют себя в том числе и на уровне общих словоформ.

Сила ассоциативной связи между любой парой терминов, по-видимому, может характеризоваться числом наборов, в которых они совместно встречаются, позиционной привязкой этих наборов (соседние или разнесенные) и, возможно, другими факторами. Вопрос требует специального изучения с привлечением экспертов ПО. Заметим также, что использование ассоциативных связей для повышения эффективности информационного поиска не всегда приводит к успеху. Тем не менее в существующих стандартах на построение тезаурусов различных предметных областей этот тип связей фигурирует.

[3] Термин «хемосорбция» формально является одногаммой, но фактически это биграмма (химическая адсорбция).

2.2. Терминологические шаблоны (образцы с переменными)

Понятие шаблона используется в различных языковых системах и подразумевает группу символьных объектов, объединенных в один класс по какому-то общему структурному признаку. Шаблоны многолики. Например, для описания регуляторных фрагментов в генетических текстах используют частично-специфицированные строки символов или строки с элементами типа “don’t care”. В качестве образца может выступать регулярное выражение. Поиск по регулярному выражению реализован, например, в системе ALEX [11]. Лексико-синтаксические шаблоны, предназначенные для распознавания специфических языковых конструкций (например, согласованных именных словосочетаний), обсуждаются в [12].

Применительно к используемой нами *L*-граммной системе представления текстов нас будут интересовать шаблоны, объединяющие в один класс сходные (отличающиеся не более, чем по 1–2 позициям) цепочки слов. Формально, такого типа шаблоны можно рассматривать как частный случай образцов с переменными (см. [13]), где переменные указывают на позиции, допускающие варьирование. Например, шаблон из 3-х слов с одной переменной *x*, имеющий форму *p = производство x кислоты*, допускает вместо *x* следующие подстановки: *серной* (встречается в исходной подборке 11 раз), *азотной* (3 раза), *пропионовой* (1), *акриловой* (1). Эти термины связаны друг с другом отношением принадлежности к одному таксону «типы кислот». Образец с двумя переменными (*x* и *y*), имеющий форму *p = окисление x в y*, допускает следующие пары согласованных подстановок: *x = этилена*, *y = этиленоксид* (эта пара встретилась 11 раз); *x = пропилена*, *y = акролеин* (16 раз); *x = метана*, *y = метанол* (1 раз) и др. Пары *x*, *y* связаны участием в одном процессе (окисление).

Связи типа «общее–частное» часто выявляются путем установления факта вложения одной *L*-граммы в другую (более длинную)[4]. Например, если в словаре имеется термин p_1 = *окисление этилена*, то поиск по образцу p_2 = *окисление этилена в x* выявит термины, являющиеся более узкими по отношению к p_1 (например, *окисление этилена в этиленоксид* или *окисление этилена в ацетальдегид*). Аналогично, сужением термина *кислотный центр* в соответствии с образцом *p = кислотный центр x* будут термины *кислотный цент Льюиса* и кислотный *центр Бренстеда*.

Формирование образцов с одной переменной осуществляется очень просто. Пусть, для примера,

[4] Все приводимые в данном абзаце примеры вложений касаются пар цепочек, относимых экспертами к терминам предметной области.

$L = 3$. Вычисляем по исходным текстам L-граммную характеристику 3-го порядка, содержащую полный спектр L-грамм, представленных в тексте, с указанием их частот. Заменяем словоформы, стоящие в первой позиции каждой 3-граммы, элементом x. При этом «склеиваются» (становятся неразличимыми) все 3-граммы, отличавшиеся только по первой позиции, и возникает множество образцов вида $p = x\ a\ в$, где a и $в$ — фиксированные канонические формы, например, $p = x$ *активированный комлекс*, где $x \in$ {*образование*, *конфигурация*, *модель*, *теория*,…}. Аналогично, заменяем элементом x словоформы, стоящие во 2-й позиции каждой 3-граммы. При этом склеиваются все триграммы, отличающиеся только по этой позиции, и возникает множество образцов вида $p = a\ x\ в$, где a и $в$ — фиксированные канонические формы, например, $p =$*образование x комплекс*, где $x \in$ {*активированный*, *поверхностный*, *мультиплетный*, *сверхкислотный*, *сульфитный*, *низкоспиновый*, *высокоспиновый*,…}. Здесь список допустимых значений содержит перечисление типов комплексов и указывает на наличие антонимических связей между ними (*низкоспиновый* → *высокоспиновый*). Наконец, осуществляя подстановку x по третьей позиции, получаем множество образцов вида $p = a\ в\ x$. Две переменные имеет смысл вводить лишь для длинных L-грамм ($L \geq 4$).

Как показывает приведенный выше пример, анализ допустимых подстановок в образцах дает важную информацию о предполагаемых связях между объектами. Возникающие при этом трудности проиллюстрируем на примере образца $p = x$ *реактор*, допускающего более 100 вариантов различных подстановок в качестве значения переменной x (на исходном материале). Среди этих подстановок можно выделить группу, характеризующую типы реакторов: *каталитический* (встретился в текстах 21 раз), *проточный* (13), *трубчатый* (9), *адиабатический* (9), *изотермический* (5) и др. Она представляет основной интерес. Другая группа подстановок характеризует конструктивные особенности реактора: *конструкция* (3), *корпус* (4), *центр* (3), *освинцованный* (1), *железный* (1). Третья группа носит характер «шума»: *промышленный* (9), *пустой* (3), *распространенный* (2), *третий* (1), *изнутри* (1), *рассчитывать* (1). Нетрудно видеть, что для выделения интересующей нас группы, важной является информация о частеречных значениях и, в меньшей степени, о частоте встречаемости. Возможность фильтрации по этим параметрам предусмотрена в программе. Но даже если устранены числительные, глаголы и наречия, а также однократно встречающиеся объекты, эксперту придется разбираться со случаями типа *промышленный* (9), *пустой* (3) и т. п.

Завершая этот раздел, заметим, что формирование образцов и их последующий анализ напоминает изучение конкордансов, собранных вместе «на все случаи жизни». Некоторая избыточность такого подхода, тем не менее, оправдана, поскольку образцы несут в себе элемент обобщения. Задавая, например, поисковый запрос в виде *окисление x в y*, мы получим не только пары *x* и *y*, фигурировавшие в исходной подборке но и многие другие в ней отсутствовавшие.

2.3. Индикаторы связи

Понятие индикатора того или иного аспекта содержания текста известно давно (см. обзор [14]). Индикаторы могут использоваться и для обнаружения в тексте упоминаний о каких-либо объектах, событиях и т. п. Применительно к интересующей нас задаче (формирование тезауруса ПО) индикаторы могут быть использованы для выявления связей между понятиями ПО. К достоинствам индикаторного подхода следует отнести простоту реализации, интерпретируемость результатов, к ограничениям — необходимость формирования индикаторных словарей в каждом отдельном случае (как правило, вручную) и отсутствие гарантий обязательного присутствия индикатора.

Иллюстрирующим примером наличия индикаторов связи в тексте может служить фраза из раздела z5 подборки, принадлежащая О.В. Крылову: «Он (Ипатьев)… создал ряд важнейших каталитических процессов нефтепереработки, *таких как* алкилирование, гидрокрекинг, изомеризация». Здесь индикатором связи «общее–частное» выступает биграмма *таких как*, роль «общего» играет выделенный левый контекст этой биграммы, а «частного» — правый.

Индикаторы связи не являются элементами терминологического словаря, но отбираются параллельно с его формированием путем просмотра экспертом устойчивых цепочек (L-грамм), упорядоченных по убыванию показателя неравномерности позиционного распределения. В отличие от терминов ПО индикатор может даже не удовлетворять требованию синтаксической связности. Таковым, например, является индикатор причинно-следственной связи *приводить к* (частота встречаемости в подборке $F = 278$).

Кроме уже упомянутых индикаторов *такой как*, *приводить к* было выделено еще несколько десятков индикаторов связи разного типа. Для их отбора эксперту пришлось просмотреть около 2000 устойчивых двухсловных сочетаний с частотой встречаемости $F \geq 10$. Напомним, что для выделения их непосредственно из текстовой подборки нужно было бы прочитать порядка 400 тыс. слов. Укажем некоторые из отобранных индикаторов: — *и др*. ($F = 284$, типы связи — «общее–частное», принадлежность к одному таксономическому классу). Текстовый пример: «В гомогенном кислотном катализе *в качестве* катализаторов используют протонные кислоты (H_2SO_4, HC_1, H_3PO_4 *и др.*)». Здесь индикатор *и др*. связывает

каждую из конкретных кислот с обобщающим термином *протонные кислоты*. Кроме этого имеется еще один индикатор *в качестве*, который связывает термины *катализатор* и *протонные кислоты*. Заметим, что индикатор *и др.* срабатывает практически без ошибок; — *один из* (F = 180, типы связи: «общее–частное», ассоциативная). Текстовый пример 1: «В нефтепереработке алкилирование используется как *один из* методов повышения октанового числа бензина». Здесь *алкилирование* ассоциативно связывается с *октановым числом*. Текстовый пример 2: «Если *один из* реагентов связывается сильно, а другой распределен равномерно между обеими фазами, то …». Этот пример иллюстрирует ситуацию, когда индикатор не срабатывает.

Кроме рассмотренных можно упомянуть такие индикаторы как *в том числе* (F = 16), *использовать в качестве* (F = 16, условная синонимия), *представлять собой* (F = 82, часто используется как элемент определения), *состоящий из* (F = 72, часть–целое), *связанный с* (F = 136) и др. Как уже было показано на текстовом примере 2, индикаторы не всегда фиксируют связь, но их значительное разнообразие и высокая частота встречаемости в тексте позволяют выявить множество связей в формируемом тезаурусе. Заметим также, что многие из перечисленных индикаторов можно трактовать и как маркеры фактов (например, *приводит к*) или элементы определений (*так называемый*, *представлять собой* и др.), что расширяет сферу применимости индикаторного подхода.

Заключение

Предложены три возможных подхода к выявлению связей между элементами тезауруса предметной области (катализ), формируемого на основе анализа достаточно представительной текстовой подборки. Рассмотрены возможности автоматизации (частичной) этого процесса в рамках используемой авторами L-граммной системы представления текстов. Они ориентированы на минимизацию неизбежного (на данный момент) ручного труда эксперта на заключительном этапе.

Литература

1. *Dobrov B., Loukachevitch N., Nevzorova O.* An approach to new ontologies development: main ideas and simulation results // Int. J. Information Theories & Applications. — Vol. 10, N 1, 2003. — P. 98–105.
2. *Гусев В. Д., Саломатина Н. В.* Алгоритм выявления устойчивых словосочетаний с учетом их вариативности (морфологической и комбинаторной) // Труды межд. конф. «Компьютерная лингвистика и интеллектуальные технологии» (Диалог–2004), М.: Наука, 2004. — С. 530–535.
3. *Гельбух А. Ф., Сидоров Г. О., Эрнандес-Рубио Э., Чубукова М. В.* Словари сочетаемости слов: какой метод составления лучше? // Там же. — С. 133–138.
4. *Браславский П. И., Соколов Е. А.* Сравнение четырех методов автоматического извлечения двухсловных терминов из текста // Труды межд. конф. «Компьютерная лингвистика и интеллектуальные технологии» (Диалог 2006), М.: Изд. РГГУ, 2006. — С. 88–94.
5. *Кузнецов П. И.* Лингвистические и алгоритмические аспекты выделения объектов и связей из предметно-ориентированных текстов // Труды межд. конф. «Компьютерная лингвистика и интеллектуальные технологии» (Диалог 2007), М.: Изд. РГГУ, 2007. — С. 333–342.
6. *Шабанов В. И., Власова А. Е.* Алгоритм формирования ассоциативных связей и его применение в поисковых системах // Труды межд. конф. «Компьютерная лингвистика и интеллектуальные технологии» (Диалог 2003), М.: Наука, 2003. — С. 603–608.
7. *Шеннон К.* Предсказание и энтропия печатного английского текста // Работы по теории информации и кибернетике. — М.: Изд. ИЛ, 1963. — С. 669–686.
8. *Гусев В. Д., Саломатина Н. В.* L-граммное представление текстов на естественном языке и его возможности // Материалы Всерос. научн. конф. «Квантитативная лингвистика: исследования и модели» (КЛИМ–2005). — Новосибирск: Изд. НГПУ, 2005. — С. 256-270.
9. *Гусев В. Д., Немытикова Л. А., Саломатина Н. В.* Выявление аномалий в распределении слов или связных цепочек символов по длине текста // Вычислительные системы, вып. 171. — Новосибирск, ИМ СО РАН, 2002. — С.51–74.
10. *Гусев В.Д, Мирошниченко Л. А., Саломатина Н. В.* Профиль кластеризуемости текстов и возможности его использования // MegaLing2006. Горизонти прикладної лінгвістики та лінгвістичних технологій. Доповіді міжнародної конференції. — Сімферополь: Вид-во «ДиАйПи», 2006. — С. 203–204.
11. *Жигалов В. А., Жигалов Д. В., Жуков А. А., Кононенко И. С., Соколова Е. Г., Толдова С. Ю.* Система ALEX как средство многоцелевой автоматизированной обработки текстов // Труды межд. конф. «Компьютерная лингвистика и интеллектуальные технологии» (Диалог–2002), М.: Наука, 2002. — С. 192–208.
12. *Большакова Е. И., Баева Н. В., Бордаченкова Е. А., Васильева Н. Э., Морозов С. С.* Лексико-синтаксические шаблоны в задачах автоматической обработки текста // Труды межд. конф. «Компьютерная лингвистика и интеллектуальные технологии» (Диалог 2007), М.: Изд. РГГУ, 2007. — С. 70–75.
13. *Handbook* of Formal Languages // G. Rosenberg, A. Salomaa (Eds), Vol.1, 1996. — Ch.4.
14. *Пащенко Н. А., Кнорина Л. В., Молчанова Т. В.* и др. Проблемы автоматизации индексирования и реферирования // Итоги науки и техники. Информатика, т. 7. — 1983 г. — С. 7–164.
15. *Остапенко В. А.* Выделение и классификация терминов с помощью элементарных квантитативных моделей // НТИ, сер. 2, № 11, 1989. — С. 24–28.

Проект типологической базы данных по синтаксическим ограничениям на передвижения

Project of the typological database on syntactic constraints on movement

Сердобольская Н. В. (serdobolskaya@gmail.com)
Российский государственный гуманитарный университет

Циммерлинг А. В. (meinmat@yahoo.com)
Московский государственный педагогический университет им. М. А. Шолохова

Аркадьев П. М. (alpgurev@gmail.com)
Институт славяноведения

Работа представляет собой пилотный проект типологической базы данных по синтаксическим передвижениям и ограничениям на передвижения. Цель настоящей базы данных — обобщение и фиксация опыта, накопленного различными исследователями в ходе изучения синтаксических ограничений в языках, принадлежащих различных языковым семьям и ареалам.

1. Описание проекта

Работа представляет собой пилотный проект типологической базы данных по порядку слов и синтаксическим ограничениям (далее: БД). Данный проект[1] реализуется группой исследователей под руководством А. В. Циммерлинга (см. сайт http://antonzimmerling.wordpress.com/current-projects/). Участники проекта: А. В. Циммерлинг, Н. В. Сердобольская, Н. Б. Пименова, П. М. Аркадьев и А. А. Перекрестенко.

БД создается как аналог типологическим базам данных, посвященных подробному описанию отдельных явлений в неродственных языках (см., например, базу данных по реципроку http://languagelink.let.uu.nl/burs/, по согласованию http://www.smg.surrey.ac.uk/Agreement/index.aspx, супплетивизму http://www.smg.surrey.ac.uk/Suppletion/index.aspx и др.). Цель настоящей БД — обобщение и фиксация опыта, накопленного различными исследователями в ходе изучения синтаксических передвижений и ограничений на передвижения в языках, принадлежащих различных языковым семьям и ареалам. В частности, использованы параметры, которые обсуждаются в работах Гринберг 1963; 1970; Dryer 1985; 1992; Hawkins 1994; Циммерлинг 2006; 2007; Franks, King 2000; Тестелец 1999; 2001. Хорошо известно, что синтаксические параметры изучены в различных языках неравномерно. В частности, в грамматиках многих языков можно найти информацию о базовом порядке слов в именной группе или о наличии *wh*-передвижения, однако лишь в немногих источниках может содержаться указание на наличие в языке синтаксических островов или последовательное перечисление ограничений на топикализацию. Мы ставим задачу разработать признаковую базу, которая может быть использована как основа для анкеты по описанию синтаксических передвижений и ограничений в произвольном языке.

Аналогичным образом, неравномерно исследовано взаимодействие различных параметров порядка слов и синтаксических передвижений и ограничений. В типологических работах по порядку слов, в частности, подробно изучены вопросы взаимодействия порядков элементов в различных типах составляющих (именные группы, предложные/послеложные группы и т. п., см. Гринберг 1970; Dryer

[1] Работа выполнена при поддержке гранта РГНФ 09-04-00297а. Авторы выражают благодарность О. И. Беляеву за помощь в технических вопросах.

1992, а также типологический проект по порядку слов http://linguistics.buffalo.edu/people/faculty/dryer/dryer/database). Однако о взаимодействии порядка слов и различных типов синтаксических передвижений на настоящий момент гораздо меньше. Планируемая БД призвана служить отправной точкой для такого рода исследований.

БД создается в формате Microsoft Access 2003. Признаковая база, лежащая в основе БД, разработана на основе параметров, обсуждаемых в (Циммерлинг 2006; 2007 и др.) и содержит следующие три группы параметров: признаки порядка слов, синтаксические ограничения и передвижения. В отличие от типологической базы данных по порядку слов, разработанной М. Драйером (http://linguistics.buffalo.edu/people/faculty/dryer/dryer/database), в настоящей базе данных сведения о базовом порядке слов даются не в качестве основного, а лишь в качестве вспомогательного параметра (необходимого для учета синтаксических ограничений на передвижения). На настоящий момент БД существует в пилотной версии, которую планируется в дальнейшем совершенствовать за счет увеличения количества признаков, уточнения и улучшения интерфейса, а также за счет привлечения материала других языков. На настоящий момент БД содержит информацию лишь о трех языках (русский, английский, даргинский). В дальнейшем планируется расширение языковой выборки за счет имеющегося у группы материала. БД реализуется на английском языке.

2. Признаковая база

Сложность построения типологической базы данных состоит в отграничении непосредственно наблюдаемых фактов от теоретических конструктов. В особенности, это касается баз данных, фиксирующих синтаксические или семантические явления. Например, хорошо известное отклонение от базового порядка слов в русском языке (см., напр., Ковтунова 1976), проиллюстрированное в (1), формально может быть описано тремя способами: как перестановка подлежащего и глагольной группы, вынос глагольной группы влево или вынос подлежащего вправо:

(31) *Кто пришел? — Пришел отец.*

Многие синтаксические явления трактуются исследователями различным образом — в зависимости от теоретической парадигмы или от традиции описания, принятой при анализе языков данного ареала или языковой семьи. Например, вынос топикального элемента в крайнюю правую позицию в романских языках (2) описывается, как Right Detachment, Right Dislocation и т. п. (см. Lambrecht 1995; Erteschik-Shir 2007; Fernandez-Vest 2005 и др.).

Французский

(2) *Je l' ai vue, ta voiture.*
я он иметь видеть твой машина
Твою машину я видел. (букв. Я ее видел, твою машину.)

В целях устранения неоднозначности, возникающей за счет использования различной терминологии, в настоящей БД делается попытка отделить непосредственно наблюдаемые факты языка от их теоретической интерпретации, принятой в литературе. Для этого все параметры делятся на два типа: параметры дескриптивного и теоретического характеры. Параметры первого типа фиксируют непосредственно наблюдаемые факты (например, базовый порядок слов в именной группе, возможность разрыва именной группы глаголом: рус. *Белую принеси миску*); параметры второго типа отражают теоретические выводы и приводятся с опорой на соответствующие теоретические статьи (например, наличие в языке локативной инверсии, топикализации и т. п.). Параметры первого типа приводятся в листаемых вкладках, параметры второго типа следуют в прямоугольнике в нижней части формы (см. рис. 1). Левая верхняя часть записи содержит базовую информацию о синтаксисе языка: название (включая сведения о генетической принадлежности), морфологический тип (агглютинативный, флективный, изолирующий, наличие инкорпорации), преобладание суффиксации или префиксации в словоизменении, базовый тип кодирования семантических ролей (аккузативный, эргативный, активный или нейтральный). Кроме того, приводится список лексических категорий языка и проблемы, возникающие при их выделении.

Перечислим используемые нами параметры. Синтаксические явления, непосредственно наблюдаемые при анализе языковых данных и требующих минимального теоретического обоснования, разделены на три группы и представлены в трех вкладках, соответственно. Используются следующие вкладки:

2.1. Базовый порядок слов и разрыв составляющих (Basic word order)

- базовый порядок слов во всех типах фразовых составляющих (например, Basic word order: NP);
- допустимость перестановок (задается в том же окошке, что и базовый порядок);
- разрывность всех типов фразовых составляющих элементами внешней составляющей (задается наличием галочки в соответствующем окошке, например, Split: Gen N; в поле комментариев содержатся указания на типы элементов, способных разрывать составляющие).

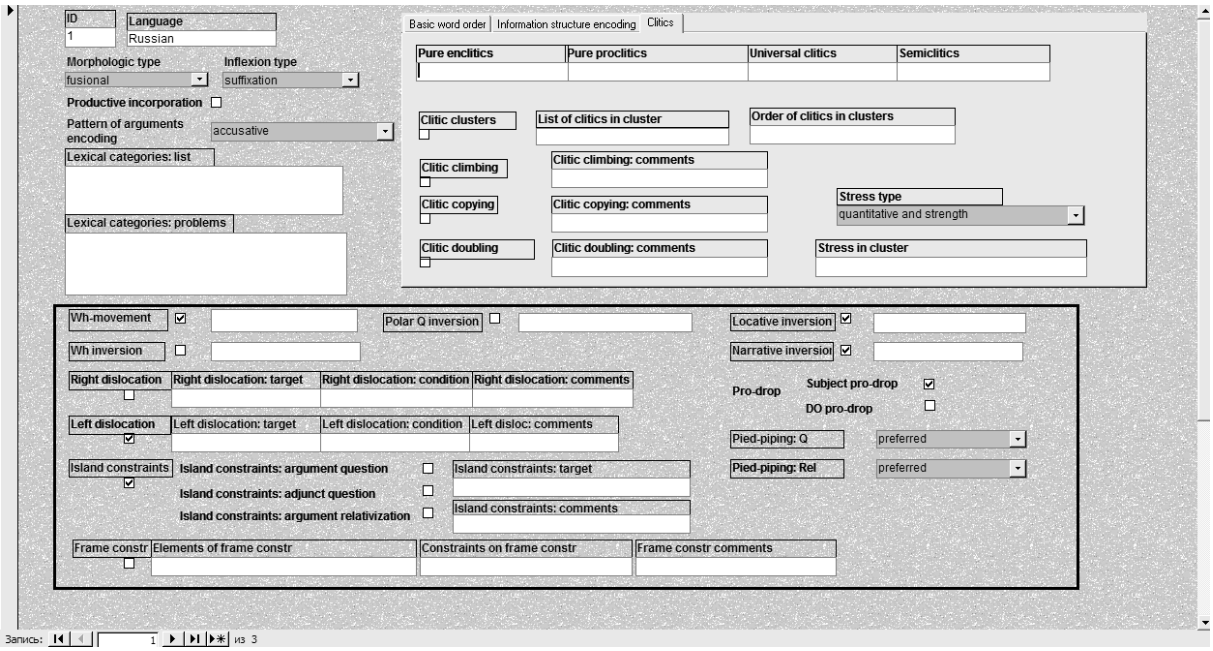


Рис. 1 База данных по порядку слов и синтаксическим ограничениям в языках мира: вкладка «клитики»

2.2. Коммуникативные категории (Information structure encoding)

- базовые средства выражения различных типов коммуникативного членения высказывания: предикатный фокус, узкий фокус, контрастивный фокус, топик и т. п. (см. Lambrecht 1995; Ertheschik-Shir 2007; Янко 2001; 2008);
- пример, глоссы и перевод на основные средства кодирования информационной структуры.

2.3. Клитики

- список клитик в анализируемом языке (если таковые имеются)
  - а) строгие энклитики (pure enclitics);
  - б) строгие проклитики (pure proclitics);
  - в) универсальные клитики (universal clitics), или клитики, способные употребляться как энклитики или проклитики;
  - г) полуклитики (semiclitics), или элементы, ведущие себя, как клитики, не во всех контекстах;
- наличие цепочек клитик (clitic clusters);
- список клитик, способных выступать в цепочке;
- порядок клитик в цепочке;
- наличие явления clitic climbing (в конструкциях с сентенциальными актантами употребление при главном глаголе клитики, семантически относящейся к зависимому глаголу, см. Rizzi 1982);
- наличие явления clitic copying (повтор клитики при одном полноударном слове);
- наличие явления clitic doubling (или местоименная реприза, т. е. употребление полной именной группы и кореферентного ей местоимения, см. (2));
- тип ударения в языке (динамическое, силовое, силовое и динамическое, музыкальное);
- взаимодействие ударения с клитиками и цепочками клитик.

Каждый параметр сопровождается полем, содержащим комментарии и ссылки на имеющуюся литературу.

В прямоугольнике в нижней части формы даны теоретические выводы о наблюдаемых синтаксических передвижениях и ограничениях. Приводимые выводы базируются на имеющихся теоретических работах либо на исследованиях, осуществляемых участниками проекта. БД содержит информацию о следующих синтаксических явлениях:

1) Pro-drop, т. е. опущение местоимений в актантных ИГ;
2) Pied-Piping, или «эффект крысолова» (необходимость передвижения вершины составляющей при передвижении зависимого);
3) *Wh*-передвижение;
4) *Wh*-инверсия (наличие инверсии в специальных вопросах);
5) *Q*-инверсия (наличие инверсии в общих вопросах);
6) Локативная инверсия;
7) Нарративная инверсия;
8) Коммуникативно нагруженные синтаксические передвижения (Right Dislocation и Left Dislocation);
9) Синтаксические острова;
10) Рамочные конструкции.

Для каждого из перечисленных параметров предусмотрено поле комментариев для более подробной информации и ссылок на имеющуюся литературу.

Параметр Pro-Drop (опущение местоимений, обозначающие актанты предиката) фиксируется отдельно для подлежащных ИГ, отдельно для ИГ, обозначающих прямое дополнение. (См. Huang 1984 относительно необходимости разграничения данных явлений.) Для фиксации явления Pied-Piping (иначе, эффект крысолова: передвижение вершины составляющей при передвижении его зависимого) предусмотрено четыре варианта заполнения: 1) Pied-Piping обязателен (напр., английский язык); 2) предпочтителен, но в большинстве случаев необязателен (напр., русский язык); 3) запрещен (авторам неизвестны такие примеры, однако они предполагаются теоретически возможными); 4) параметр неприменим, т. к. в языке нет синтаксических передвижений на поверхностном уровне (т. н. in situ языки; по-видимому, к таким языкам относится китайский). В БД отдельным образом отражается информация о явлении Pied-Piping при частном вопросе и релятивизации.

Коммуникативно нагруженные синтаксические передвижения (параметры Right Dislocation и Left Dislocation) объединяют все передвижения, кодирующие различные коммуникативные категории — включая топикализацию, передвижение фокуса влево, передвижение фокуса вправо, передвижение топика вправо (Right Detachment, ср. (2)) и т. п. (см. Циммерлинг 2007). Такое объединение перечисленных типов передвижений обосновано следующим. Не вызывает сомнений, что, например, различные типы передвижения фокусных составляющих влево существенным образом различаются в языках мира. Поэтому было бы неправомерно объединять их под одним терминам. С другой стороны, в силу традиции определенный тип явлений тем не менее принято объединять одним термином — речь идет о топикализации. Несмотря на очевидные различия в морфосинтаксическом и интонационном оформлении топика и ограничениях на топикализацию в разноструктурных языках, термин «топикализация» активно используется различных лингвистических парадигмах и описательных традициях. Такое объединение мотивировано сходством описываемых явлений в разных языках.

Таким образом, используемая исследователями терминология не дает равномерного описания для различных коммуникативно нагруженных типов передвижений: некоторые типы передвижений объединяются в большие группы, а некоторые получают отдельные наименования. Данные наименования в большинстве случаев не являются общепринятыми, а в значительной мере зависят от теоретической парадигмы исследователя, описывающего язык. При сохранении терминологии в настоящей БД пришлось бы тем не менее для каждого случая приводить подробные пояснения относительно конкретного типа передвижения и ограничений на его использования в данном конкретном языке. Учитывая данные соображения, в рамках данной БД было решено полностью отказаться от принятой терминологии, разделяя все имеющиеся в языке передвижения на две основные группы: передвижения вправо и передвижения влево. Далее, созданы поля для уточнения, о каком типе передвижения идет речь: поле Right/Left Dislocation target служит для указания на коммуникативную категорию, вызывающую передвижение (топик, фокус, узкий фокус, контрастивный топик и т. п.), поле Right/Left Dislocation condition содержит информацию об ограничениях на данное передвижение (например, в даргинском языке очень редко допускается передвижение генитива из именной группы и т. п.). Наконец, поле Right/Left Dislocation comments предназначено для остальных комментариев и ссылок на имеющуюся литературу. Таким образом, отказ от использования терминов, принятых для различных типов передвижений, позволяет выявить сходства и различия в системах топикальных и фокусных передвижений различных языков, что необходимо для систематизации имеющейся информации и унификации терминологии в данной области.

Параметр «Синтаксические острова» включает в себя следующую информацию. Фиксируется наличие следующих типов островных ограничений в языке: ограничения на вопрос и релятивизацию из различных типов составляющих. В поле Island constraints: target перечисляются те типы составляющих, для которых действуют данные ограничения. Поле Island constraints: comments содержит комментарии и ссылки на имеющуюся литературу.

Рамочные конструкции характеризуются по следующим признакам. Поле Elements of frame constructions содержит указание на элементы, способные образовывать рамочную конструкцию. В поле Constraints on frame constructions приводятся данные об условиях, при которых реализуется рамочная конструкция. Данные сведения снабжены комментариями и ссылками на источники (поле Frame constructions comments).

Литература

1. *Dryer M. S.* Greenbergian word order correlations // Language. 1992. Vol. 68. N 1. 81–138.
2. *Dryer M. S.* Word Order // T. Shopen (ed.) Language Typology and Syntactic Description, vol. 1: Clause Structure. Cambridge University Press. Cambridge: 1985.
3. *Erteschik-Shir N.* Information structure. The syntax — discourse interface // Oxford University Press: 2007.
4. *Fernandez-Vest J.* Vers une typologie linguistique du détachement à fondement ouralien d'Europe // Bulletin de la Société de Linguistique de Paris, BSL, C1, 1. Paris: 2006. Pp. 173–224.
5. *Franks S., King T. H.* A Handbook of Slavic clitics. Oxford University Press. Oxford: 2000.
6. *Hawkins J. A.* A performance theory of order and constituency // Cambridge University Press. Cambridge: 1994.
7. *Huang C., James T.* On the Distribution and Reference of Empty Pronouns // Linguistic Inquiry 15: 1984. Pp. 531–574.
8. *Lambrecht K.* Information Structure and Sentence Form. Topic, Focus and the Mental Representation of Discourse Referents // Cambridge University Press: 1995.
9. *Rizzi L.* Issues in Italian Syntax // Foris: 1982.
10. *Ross J. R.* Constraints on Variables in Syntax. Ph.D.thesis // Massachusetts Institute of Technology, Cambridge: 1967.
11. *Гринберг Дж.* Квантитативный подход к морфологической типологии языков // Новое в лингвистике // М., 1963. Вып. III. С. 60–95.
12. *Гринберг Дж.* Некоторые грамматические универсалии, преимущественно касающиеся порядка значимых элементов // Новое в лингвистике // М.: 1970. Вып. V. С. 114–162.
13. *Гринберг Дж., Осгуд Ч., Дженкинс Дж.* Меморандум о языковых универсалиях // Новое в лингвистике // М.: 1970. Вып. V. С. 31–44.
14. *Ковтунова И. И.* Современный русский язык. Порядок слов и актуальное членение предложения // Просвещение. М.: 1976.
15. *Тестелец Я. Г.* Порядок слов и структура составляющих // А. Е. Кибрик (ред.) Элементы цахурского языка в типологическом освещении // «Наследие». М.: 1999. С. 293–346.
16. *Тестелец Я. Г.* Введение в общий синтаксис // РГГУ. М.: 2001.
17. *Циммерлинг А. В.* Отношение свободного порядка слов и модели инверсии // Труды международного семинара «Диалог 2006: компьютерная лингвистика и информационные технологии» // М.: 2006. С. 541–544.
18. *Циммерлинг А. В.* Порядок слов в русском языке // Текст, Структура и семантика. Доклады XI международной конференции // МГГУ. М.: 2007. С. 138–151.
19. *Янко Т. Е.* Коммуникативные стратегии русской речи. М.: 2001.
20. *Янко Т. Е.* Интонационные стратегии русской речи в сопоставительном аспекте. М.: 2008.

Язык описания правил в системе лексического анализа Ея-текстов DICTASCOPE TOKENIZER

The language for describing rules in DICTASCOPE TOKENIZER — a system for lexical analysis of natural language texts

Скатов Д. С. (ds@dictum.ru), **Ливерко С. В.** (sl@dictum.ru),
Вдовина Н. А. (vn@dictum.ru), **Окатьев В. В.** (oka@dictum.ru)

ООО «Диктум», Нижний Новгород, Россия

В статье обсуждается лексический анализ ЕЯ-текстов в контексте задач извлечения информации (именованных сущностей, гипертекстовых переходов, терминологии) на основе правил. Язык DSTL, предлагаемый для их решения, обладает разнообразными выразительными средствами при высокой компактности описаний.

1. Введение

Решения для извлечения структурированных данных из неразмеченных ЕЯ-текстов востребованы во множестве приложений. Можно строить прогнозы, автоматически извлекая из новостных потоков упоминания персон, организаций, адресов, торговых марок и отношений между ними с учетом временно́й динамики; автоматизировать построение баз данных, находя в текстах наименования, коды, характеристики товаров; автоматически добавлять к существующим веб-страницам семантическую разметку.

Для решения всех этих задач традиционно используются механизмы применения *правил*, составляемых экспертом (возможно, в полуавтоматическом режиме). По правилу механизм определяет в тексте фрагмент с данными и извлекает их в определенном формате. База правил должна допускать пополнение и поддержку с адекватными трудозатратами. Для практического использования к механизму предъявляются требования по скорости: недопустим комбинаторный рост времени анализа при линейном росте количества правил.

В данной работе представлен язык описания правил DSTL (DictaScope Tokenizer Language), используемый в системе лексического анализа ЕЯ-текстов DictaScope Tokenizer (DST, от англ. *to tokenize* — снабжать метками, помечать). Он разрабатывался для создания быстрого и относительно простого в применении механизма лексического анализа, включающего функции уже известных разработок. Средства, находящиеся в открытом доступе [3, 12], не позволяли достичь требуемых свойств механизма, поэтому язык создавался независимо от них. Был учтен опыт специалистов, занимающихся решением схожих задач, и результаты собственных исследований [13].

2. Постановка задачи

Любая задача, затрагиваемая в данной работе, может быть сформулирована следующим образом: выявить в *неразмеченном* ЕЯ-тексте *лексические конструкции* — цепочки *слов* входного текста (*возможно, разрывные*), каждая из которых снабжается *набором данных* определенной структуры. Структура включает:

- имя *класса*, которому принадлежит конструкция;
- *нормальную форму конструкции*, которая состоит из *нормализованного текстового представления* (удобного для прочтения человеком) и набора именованных полей с присвоенными значениями.

Эта задача далее называется *лексическим анализом* [15] *естественного языка*, или, сокращенно — *LANL* (Lexical analysis of natural language, [1]).

В данной постановке может быть сформулирована известная задача *извлечения именованных сущностей* (NER — *Named Entity Recognition*, [5, 6, 9, 11]), к которым относят имена персон, организации, географические адреса, даты, результаты измерений

и пр. Для таких сущностей характерна вычислимая нормальная форма, а упоминания об одной сущности записываются в тексте в разных формах и являются *непрерывными* символьными конструкциями. В качестве классов выступают типы сущностей («*Дата*», «*Персона*», «*Организация*»), а нормальная форма — это набор полей. Напр., запись «***31 июля 1986 г.***» (равно как и записи «***1986/31/7***», «***July 31st, 1986***») естественно отнести к классу «*Дата*» и снабдить нормальной формой в виде трех числовых полей *{Day = 31, Month = 7, Year = 1986}* и текстового представления «***31.07.1986***».

В материалах о приложении текстовых шаблонов к обработке ЕЯ эти шаблоны обозначаются поразному: встречаются лексические [4], семантические [6], лексико-синтаксические [12] шаблоны. Задачу NER часто относят к семантическому анализу [14]. Пересекаются понятия лексического анализа, токенизации, графематики. Для ясности изложения авторы далее уточняют интерпретацию этих терминов.

В LANL конструкции формируются из слов. Задачу разбиения текста на слова называют (символьной) токенизацией [19]. В теории формальных языков токенизация и лексический анализ — синонимы [15]. В обработке ЕЯ это, вообще говоря, не так. В публикациях по теме: (1) токенизация — это разбиение цепочек символов на слова [19], (2) лексический анализ — формирование конструкций из цепочек слов [1, 4]; выход (1) является входом (2). Решение этой задачи представляет самостоятельный интерес: она сложна для арабского, японского языков, любого текста с опечатками типа пропуска пробелов. В статье эта задача детально не рассматривается. Модель слова, принятая в языке DSTL, допускает адаптацию к любой схеме символьной токенизации.

При извлечении конструкций полезно выявить *отношения* между ними — такова основная цель *семантического анализа текста*. Эти отношения могут быть описаны лексическими конструкциями, которые, в отличие от именованных сущностей, чаще всего разрывны. Напр., *«Персона»* могла в прошлом являться сотрудником *«Организации»* — это образует *факт «Место работы»*: «***Василий Петров**, мечтая о научной карьере, долгое время успешно **трудился в НИИ ЧАВО***». Существенные фрагменты выделены жирным, они образуют утверждение вида «*X трудился в Y*», а деепричастный оборот и уточнения для указанного факта несущественны.

Подходы, применяемые для выявления отношений и по сути решающие задачу LANL, используют элементы синтаксического анализа ЕЯ:

- применяется доступ к грамматическим значениям,
- явно присутствует дерево непосредственных составляющих (напр.: «*[[19 января 2010 г.]*$_{\text{Дата}}$ *вступил в должность [Заместитель [председателя [правительства РФ]* $_{\text{Организация}}$ *]* $_{\text{Должность}}$ *[Хлопонин А. Г.]*$_{\text{Персона}}$ *]*$_{\text{Должность}}$*]* $_{\text{Назначение}}$»).

Для выявления отношений требуется покрыть конструкциями лишь отдельные фрагменты текста; современный синтаксический анализ решает более широкую задачу — построение полного дерева предложения [13], по этой причине сложность его на порядки выше. Чтобы избежать оговорок, задачу извлечения отношений *на основе шаблонов* авторы работы также относят к LANL, без какого-либо упоминания о синтаксисе.

Полное представление о роли LANL в решении задач анализа ЕЯ дано на Рис. 1.

Следующие задачи, отличные от NER, также сводятся к LANL:

Отыскание в неразмеченном документе фрагментов, содержащих ссылки на другие документы и собственные разделы [10]: «*согласно ст.15 Конституции РФ*»;

Выявление в ЕЯ-текстах фраз-определений авторских терминов, их синонимов и связанных атрибутов [12]: «*Синтаксический анализ — это …*»;

Нормализация слабоструктурированных источников данных: автоматизированное формирование и коррекция номенклатурных списков (имущества, оборудования и т. д.) [4];

Т. н. прошивка законодательства — извлечение инструкций (связанных с обновлением текстов во времени) для их последующего применения [10]: «*Часть первую статьи 41 дополнить словами “или его заместителем”*».

Графематический анализ: к нему относят выявление в тексте простых лексических конструкций (ФИО с инициалами, электронные адреса, имена файлов), а также предложений, абзацев, заголовков, примечаний [7].

Анализ искусственных языков, как правило, осуществляется механизмами на основе формальных грамматик [14]. Их возможности в задачах анализа ЕЯ ограничены. В то же время, в рамках LANL они остаются эффективным описательным средством при условии соответствующих доработок, которые и были выполнены рядом исследователей [1–4, 6, 8–12, 17].

3. Обзор

Все известные авторам средства, решающие задач LANL с помощью шаблонов, основаны на регулярных выражениях. В 90-х был предложен инструмент NLlex [1], расширяющий Unix-утилиту lex доступом к морфологическому словарю и лингвистическим деталям текста. Тогда же был предложен язык CPSL [2]: в нем шаблон представляет собой решающее правило, в левой части которого записано регулярное выражение относительно свойств слов, в правой — действие, исполняемое при его срабатывании. Шаблоны CPSL можно наследовать. Этот

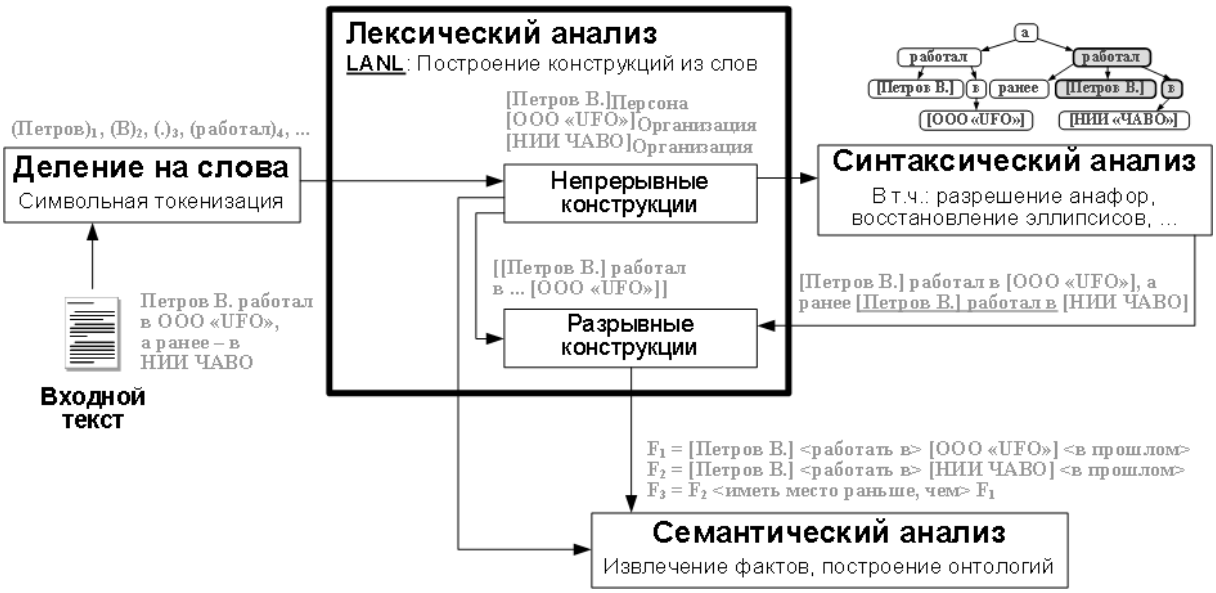


Рис. 1. Роль задачи LANL в общей схеме анализа ЕЯ

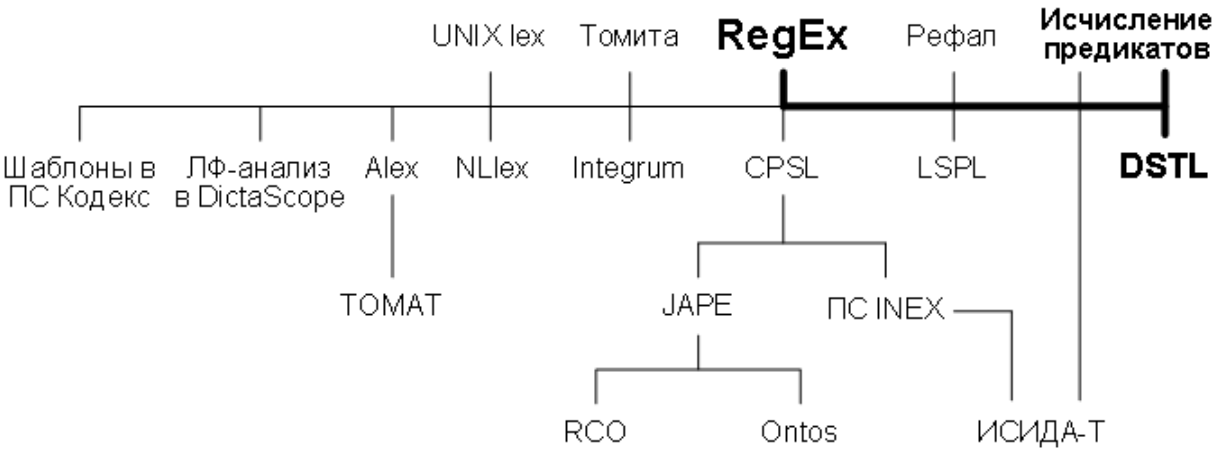


Рис. 2. «Генеалогическое древо» языковых средств для описания лексических шаблонов

язык перенял недостаток обычных регулярных выражений (regex-ов): их восприятие затруднено из-за большого числа парных скобок, а в CPSL это усугубляется наличием в шаблоне разных типов скобок. В языке JAPE [3] к CPSL была добавлена атрибутно-объектная модель текста для применения в рамках языка JAVA. Ни CPSL, ни JAPE не предусматривают проверок синтаксического согласования.

Для анализа русского языка на базе CPSL был создан язык в ПС INEX [8], на базе JAPE — языки в системах RCO Pattern Extractor [6] (ориентирован на извлечение именованных сущностей) и Ontos-Miner [9] (анализ новостных текстов). В системе «ИСИДА-Т» [17] возможности CPSL улучшены: расширены типы данных, можно создавать поля с произвольными значениями, формулировать проверки в виде логических предикатов (в частности — проверять согласование).

Ряд отечественных разработок вне ветви развития CPSL предназначен для решения частных задач: напр., ИПС «Кодекс» [10], Alex [4] (продолжила развитие в системе TOMAT). Функция т. н. лексико-фрагментационного анализа (разбиение текста на предложения с выявлением регулярных элементов типа дат и URL) встроена в синтаксический анализатор ЕЯ DictaScope [13]. Все это суть расширения regex-ов, но, в отличие от NLlex, они не учитывают морфологию.

За рамками статьи оставлено решение задач LANL методами машинного обучения. Этот подход широко применяется зарубежными исследователями в задаче NER [5], а в России представлен разработками IXLab [18]. Также можно отметить подход системы Integrum [11] (выявление именованных сущностей и их отношений), использующий алгоритм Томиты.

В работе [12] предложен язык LSPL с компактными лексическими шаблонами с наследованием и проверкой согласования. Он был применен создателями для выявления терминов в научной литературе, т. о. эти шаблоны изначально ориентированы на ограниченный тип конструкций [16]. Важные функции, свойственные CPSL, напр., работа с вычислимыми полями, в LSPL отсутствуют. В нем используются элементы, заимствованные из функционального программирования, поэтому синтаксис LSPL весьма необычен.

В силу требований гибкости и скорости ни одна из названных разработок не была положена в основу языка DSTL. Он построен на основе регулярных выражений и формализма исчисления предикатов. DSTL идейно и функционально близок к ИСИДА-Т (наиболее «продвинутому» потомку CPSL) и языку LSPL (при создании которого, согласно [12], принимались во внимание практически все языки из обзора). Можно утверждать, что DSTL покрывает возможности этих разработок.

4. Язык DSTL

4.1. Лексические правила

Словом в DSTL по умолчанию является цифробуквенное сочетание (*«январь»*, *«25»*), одиночный разделитель (*«.»*, *«–»*) или подряд идущие пробельные символы. *Атом* — это цепочка из одного и более таких слов, первое и последнее из которых — непробельные: *«январь»*, *«несмотря на то что»* — атомы, *« 23 »* — не атом. Эта схема позволяет эксперту не беспокоиться о таких деталях, как наличие пробелов между словами и их количество (DSTL позволяет учесть их при необходимости).

Правило DSTL образовано несколькими секциями. Напр., правило с именем `Year` описывает конструкции вида *«1986 г.»*:

```
Year { /* 1986 г. { year = 1986; } */
   T := Y "г." ?;
   C := Length (Y) = 4 & IsNumeric (Y);
   A := { year := Y; };
};
```

Шаблон (секция с именем **T** — *template*) — это регулярное выражение, записанное относительно его *элементов* — имен атомов и унаследованных конструкций. В шаблоне правила `Year` утверждается: конструкция состоит из атома `Y` и двухсловного атома `"г."`, который может и отсутствовать (согласно оператору `?`).

Свойства `Y` формулируются в *критерии* (секция **C** — *criterion*). По сути — это предикат, составленный из доступных в DSTL функций с аргументами-элементами шаблона **T**. Доступны логические, арифметические, строковые функции. В секции **C** примера сказано: (1) атом `Y` состоит из четырех символов (`Length (Y) = 4`), (2) `Y` — число (`IsNumeric (Y)`).

Действие (секция **A** — *action*) выполняется, если критерий сработал (т. е. оказался истинным). Оно представляет собой последовательность операций. В действии примера строковая запись года присваивается полю `year`.

4.2. Схема работы с морфологией

В DST *грамматическое значение* (*ГЗ*) представляет собой набор *грамматических характеристик* единицы языка. К ним относятся грамматические категории (часть речи, род, число, падеж, время, …) и начальная форма слова. DST не снимает омонимию — в нем слово *«для»* будет иметь два ГЗ, соответствующие глаголу *«длить»* и предлогу *«для»*. Поэтому для единицы языка приходится оперировать *множеством* ГЗ. Лексическим конструкциям также присваивается ГЗ, поэтому на согласование можно проверить не только отдельные атомы, но и (1) атомы и конструкции, (2) пары конструкций.

Проверить наличие ГЗ с заданными характеристиками можно функцией `HasGrammarForm`. При этом строится т. н. *сужение множества ГЗ*: в рамках текущей конструкции остаются только ГЗ элемента с заданными характеристиками. Напр., множество ГЗ слова `W = "Александра"` образовано тремя элементами; они имеют одинаковые значения типа, подтипа и числа `{Type: Noun, Subtype: Name, Number: Sg}`, но различаются по роду и падежу: `W.GrV = [{Case: Nom, Gender: Fem}` (ж. р. им. п.), `{Case: Gen, Gender: Masc}` (м. р. род. п.), `{Case: Acc, Gender: Masc}` (м. р. вин. п.)]. Если описываются конструкции женских имен, нужно задать проверку `F(W) = HasGrammarForm (Name, {Type: Noun, Subtype: Name, Gender: Fem})`. Тогда, `F(W) = true` и `W.GrV = [{Case: Nom, Gender: Fem}]` (ГЗ с мужским родом оказываются отфильтрованными).

Два (и более) элемента можно проверить на согласование — этой цели служит функция `AreConcordant`. Она находит среди ГЗ первого и второго элемента пару значений, у которых совпадают заданные характеристики. Напр., если `V = "студентка"`, `W = "Александра"`, то `AreConcordant (V, W, {Number, Gender, Case})` проверит согласование `V` и `W` по числу, роду и падежу, результатом будет `true`, а при `V = "автомобиль"` результат — `false`. Если ранее функция `HasGrammarForm` построила сужение для `V` или/и `W`, то при проверке будут учитываться только ГЗ из этих сужений.

Схема, принятая в DSTL для обработки ГЗ, показана на Рис. 3. Фактически, она дает способ снятия омонимии по контексту. На последующих этапах обработки конструкций (напр., синтаксическим анализатором) для них будет известно конкретное ГЗ, выведенное лексическими правилами языка DSTL.

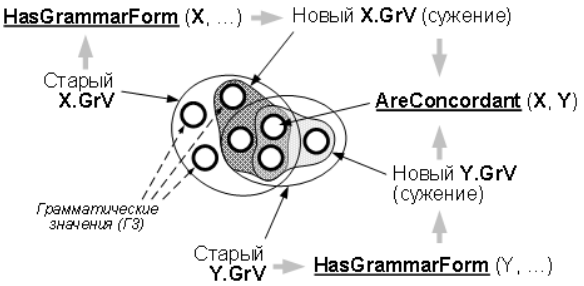


Рис. 3. Схема работы с грамматическими значениями в DSTL

4.3. Наследование конструкций

По структуре база правил в DSTL представляет собой КС-грамматику [15], в которой терминалы — это атомы, нетерминалы — имена правил. Операция использования одной конструкции в другой называется *наследованием*. Она подразумевает не только подстановку шаблона, но также и возможность обращаться к полям вложенной конструкции и наследовать их.

Далее правило `Year` (см. выше) снабжается дополнительными правилами для выделения дат: результирующая база описывает конструкции вида *«31 июля 1986 г.»*.

```
Months := { "января": 1, … ,"декабря": 12 };

Day {
   T := D; /* 31 {day: 31} */
   C := IsNumeric (D) & DiapStr (D, 1, 31);
   A := { day := StrToInt (D); };
};

Month {
   T := M; /* июль, июля {month: 7} */
   C := M in Months; /* Months["июля"] = 7 */
   A := { month := Months[M]; };
};

/* 31 июля {day: 31, month: 7} */
Day_Month { T := [Day] [Month]; };
/* 31 июля 1986 г. {day: 31, month: 7, year: 1986} */
Date { T := [Day_Month] [Year]; };
```

В строке *«31 июля 1986 г.»* можно определить три значащие конструкции: в силу правил `Day_Month`, `Year` и `Date`. Т. к. `Date` наследует `Day_Month` и `Year`, результатом анализа является *дерево* с этими конструкциями в вершинах. Набор таких деревьев, узлами которых являются все конструкции текста, называется *лексическим покрытием* (уточнение понятия *аннотаций* [17]).

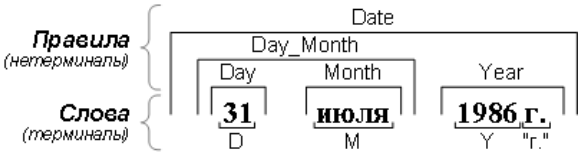


Рис. 4. Лексическое покрытие текста конструкциями класса «Дата»

При наследовании возможны конфликты конструкций, напр.: (1) *«Александра Иванова»* может быть персоной женского рода в им. п., а может — мужского в род. п., (2) во фрагменте *«Пушкин А. С. Поэмы»* могут быть определены персоны *«Пушкин А. С.»* и *«А. С. Поэмы»*, (3) для *«заместителя управляющего делами президента РФ Павла Бородина»* возможны три варианта покрытия (см. Рис. 5).

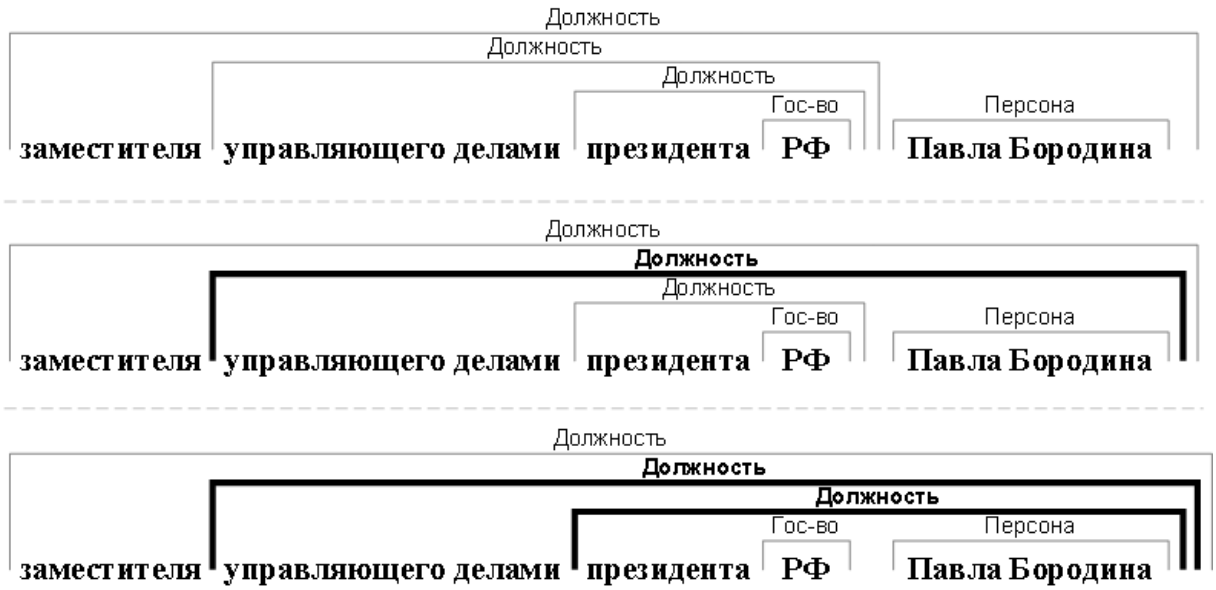


Рис. 5. Три различных дерева, соответствующие одной конструкции

Для разрешения конфликтов в DST существует система *вычислимых* весов, но эта тема выходит за рамки статьи.

4.4. Согласование и нормальная форма

Следующий пример позволяет извлечь упоминания персон мужского рода (ед. и мн. ч.) вида *«Петровым Сергеем Николаевичем»*, *«Иваны Петровы»* и т. д. Определяется правило `N` для имени:

```
N {
   T := W; /* Иван, Петру, Сергеем */
   C := HasGrammarForm (W, {Subtype: Name, Gender: Masc});
   A := { GrV := W.GrV; W := GetInitialForm (W); };
};
```

Затем аналогично `N` строятся правила `Sn` (фамилия) и `Pt` (отчество) и описываются персоны:

```
N_Sn { T := [N] [Sn]; /* Иванам Петровым */
   C := AreConcordant (N, Sn, {Gender, Number, Case}); };
Sn_N { T := [Sn] [N]; /* Петрову Ивану */
   C := AreConcordant (N, Sn, {Gender, Number, Case}); };
N_Pt_Sn { T := [N] [Pt] [Sn]; /* Ивана Михайловича Петрова */
   C := AreConcordant (N, Pt, Sn, {Gender, Number, Case}); };
Sn_N_Pt { T := [Sn] [N] [Pt]; /* Петровым Иваном Михайловичем */
   C := AreConcordant (Sn, N, Pt, {Gender, Number, Case}); };
```

Т. к. проверяется согласование, то в тексте *«Иван Петрова не видел»* персоны *«Иван Петрова»* (`N_Sn`) найдено не будет, но найдутся *«Иван»* (`N`) и *«Петрова»* (`Sn`).

4.5. Сравнение языков

CPSL был выбран для сравнения как наиболее «плодовитый» представитель шаблонных языковых средств (см. Рис. 2), близкий по функциональности к DSTL. Средства без поддержки морфологии [1, 4] по возможностям далеки от DSTL. LSPL, несмотря на идейную близость к DSTL, представляет другой крайний случай — все его существенные возможности сосредоточены исключительно в морфологии. Следующие примеры взяты в [8].

В этом сравнении, по мнению авторов, прослеживается бóльшая выразительность DSTL.

5. Алгоритм анализа

Далее дается краткое описание подхода к анализу, который применяется в DST.

Пусть $A = \{a_1, \ldots, a_N\}$ — набор атрибутов, таких, что для произвольного слова V $a_k(V) \in \{0,1\}$. Для каждого слова W, которое должно попасть в конструкцию, определяемую правилом R, из шаблона и критерия выявляется множество атрибутов $\mathsf{a}(W) \subseteq A$, значения которых для этого слова W должны быть истинными (значения остальных атрибутов для W не важны).

В шаблон вместо имени любого его слова W подставляется множество $\mathsf{a}(W)$. Далее полученный шаблон T трактуется как обычное регулярное выражение, в котором вместо символов рассматриваются множества $\mathsf{a}(W)$, вместо отношения равенства двух символов (W из шаблона и V из проверяемого текста S) — отношение вхождения подмножества во множество: $W \subseteq V$.

Пример. Пусть $A = \{1,2,3\}$, $T = \{1,2\}\ \{3\}$, $S = \{1,2\}\ \{1,2,3\}\ \{2,3\}\ \{3\}$. Шаблон T входит в текст S в смещении 1 (т. к. $\{1,2\} \subseteq \{1,2\}$, $\{3\} \subseteq \{1,2,3\}$) и 2 (т. к. $\{1,2\} \subseteq \{1,2,3\}$, $\{3\} \subseteq \{2,3\}$), но не в 3 (т. к. $\{1,2\} \not\subseteq \{2,3\}$).

Пример 1: Извлечение IP-адресов вида *«192.168.1.72»*.

CPSL

```
Rule: IPAddress (
   (Token.kind == number) (Token.string == ".")
   (Token.kind == number) (Token.string == ".")
   (Token.kind == number) (Token.string == ".")
   (Token.kind == number))
: ipAddress --> :ipAddress.Address = {
     kind = "ipAddress"}
```

DSTL

```
IPAddress {
   T := N ("." N) (3);
   C := IsNumeric (N);
   A := { kind := "ipAddress" };
};
```

Пример 2: Извлечение имен печатных СМИ: *«Газета “Веселый огородник”»*, *«Журнал “Smog”»*.

CPSL

```
Macro: NOT_QUOTE ({!Token.string == "\""})
Rule: NewspaperName
   ({Token.string =| "газета"} | {Token.string =| "журнал"})
   {Token.string == "\""}
   (({!Token.string == "\"", Morpho.Capitalized == True})
    NOT_QUOTE? NOT_QUOTE? NOT_QUOTE?)
    : newspaperName {Token.string == "\""}
--> :newspaperName.ProperName = {
   kind = "Newspaper", rule = "NewspaperName"}
```

DSTL

```
QUOTE := "\"";

Name : hidden {
   T := First (Other) (0,3);
   C := IsCapitalized (First) &
        First != QUOTE & Other != QUOTE;
};

Newspaper {
   T := Pr QUOTE [Name] QUOTE;
   C := Pr %in {"газета", "журнал"};
   A := { kind := "Newspaper"; newspaperName := Name; };
};
```

6. Заключение

Язык DSTL представляет собой новую ветвь развития современных средств описания ЕЯ-шаблонов. Разработчики учитывали достоинства и недостатки существующих разработок, чтобы создать емкий и выразительный язык, который сохранил бы полезную функциональность этих средств, но освободил составителя решающих правил от известных проблем. К ним относятся высокая сложность вложенных конструкций в регулярных выражениях и необходимость учета лишних деталей, таких как особенности объектной модели текста [3] или наличие пробелов между словами [1].

Признано, что регулярные выражения обладают большой выразительностью, но их экземпляры из реальной жизни нельзя назвать понятными. Назначение языка DSTL — устранить сложность regex-ов в задачах анализа ЕЯ, сделав описания достаточно короткими и естественными, отражающими лингвистическую суть конструкций. Описание структуры шаблона отделено от фиксации свойств его элементов, что, по мнению авторов, предотвращает разрастание шаблонов и размытие их смысла и, тем самым, упрощает расширение и поддержку правил.

Язык DSTL используется авторами для решения задач LANL, которые были обозначены в начале статьи. На данный момент можно утверждать, что в плоскости извлечения именованных сущностей язык достиг устойчивой фазы: имеются достаточно качественно функционирующие описания классов *«Персона»*, *«Организация»*, *«Время и дата»* и пр. Напр., извлечение объектов типа *«Персона»* выполняется на машине Athlon 3,1 GHz в объеме памяти 50 Мб со следующими усредненными по корпусу текстов из WikiPedia показателями: скорость 700 Кб/с, точность ~ 0,8; полнота ~ 0,9.

Ведутся исследования по применению DSTL для извлечения фактов на основе разрывных конструкций для извлечения фактов. В ближайших планах — исследования по автоматизированному составлению решающих правил на основе методов машинного обучения.

Литература

1. *Almeida J. J.* Nllex — a tool to generate lexical analysers for natural language. Technical Report UMDI95.04 // Universidade do Minho, Departamento de Informatica: 1995.
2. *Appelt D., Onyshkevych B.* The Common Pattern Specification Language // Annual Meeting of the ACL. Proceedings of a workshop on held at Baltimore, Maryland, October 13–15. Association for Computational Linguistics, Morristown, NJ, USA: 1998. P. 23–30.
3. *Cunningham H., Maynard D., Tablan V.* JAPE: A Java Annotations Pattern Engine. Technical Report CS-00-10 // University of Sheffield. UK: 2000.
4. *Жигалов В. А., Жигалов Д. В., Жуков А. А., Кононенко И. С., Соколова Е. Г., Толдова С. Ю.* Система Alex как средство для многоцелевой автоматизированной обработки текстов // Труды международного семинара Диалог-2002. М.: Наука, 2002. Т. 2. С. 192–208.
5. *Bender O., Och F. J., Ney H.* Maximum Entropy Models for Named Entity Recognition // CoNLL-2003, 7th Conference on Computational Natural Language Learning. Edmonton, Canada: May 2003. P. 148–152.
6. *Ермаков А. Е., Плешко В. В., Митюнин В. А.* RCO Pattern Extractor: компонент выделения особых объектов в тексте // Информатизация и информационная безопасность правоохранительных органов: XI Международная научная конференция. Сборник трудов. М.: 2003.
7. *Ножов И. М.* Морфологическая и синтаксическая обработка текста (модели и программы) // Интернет-публикация авторской диссертации на сайте www.aot.ru. М.: 2003.
8. *Программная* система извлечения информации из текстов (ПС INEX). Программная документация // 04832915.10028-01 33 01. М.: 2004.
9. *Дудчук Ф. И., Шафирин А. Ю.* Извлечение информации из франкоязычных текстов: морфология, синтаксис, объекты // Девятая Национальная конференция по искусственному интеллекту с международным участием КИИ-2004: Труды конференции. В 3-х т. Т. 2. М.: Физматлит, 2004. С. 489–497.
10. *Губин М. В., Меркулов А. И.* Автоматическое выделение гипертекстовых переходов в текстах документов // Труды международной конференции «Диалог-2004». М.: Наука, 2004. С. 155–158.
11. *Гершензон Л. М., Ножов И. М., Панкратов Д. В.* Система извлечения и поиска структурированной информации из больших текстовых массивов СМИ. Архитектурные и лингвистические особенности // Труды международной конференции «Диалог-2005». М.: Наука, 2005.
12. *Большакова Е. И., Баева Н. В., Бордаченкова Е. А., Васильева Н. Э., Морозов С. С.* Лексико-синтаксические шаблоны в задачах автоматической обработки текста // Труды международной конференции «Диалог 2007». М.: Изд-во РГГУ, 2007. С. 70–76.
13. *Окатьев В. В., Гергель В. П., Алексеев В. Е., Таланов В. А., Баркалов К. А., Скатов Д. С., Ерехинская Т. Н., Котов А. Е., Титова А. С.* Отчет о выполнении НИОКР по теме: «Разработка пилотной версии системы синтаксического анализа русского языка» (инвентарный номер ВНТИЦ 02200803750) // М.: ВНИТЦ, 2008.
14. *Азарова И. В., Гребеньков А. С., Ландо Т. М.* Использование маркеров актантных позиций при анализе деловых текстов для расширения логической схемы предметной области // Труды международной конференции «Диалог 2008». М.: РГГУ, 2008. С. 11–17.
15. *Ахо А. В., Лам М. С., Сети Р., Ульман Д. Д.* Компиляторы: принципы, технологии и инструментарий, 2 изд. М.: «Вильямс», 2008.
16. *Рабчевский Е., Булатова Г., Шарафутдинов И.* Формализм записи лексико-синтаксических шаблонов в задаче автоматизации процесса построения онтологий // Труды десятой всероссийской научной конференции «RCDL'2008». Дубна: ОИЯИ, 2008. С. 415.
17. *Кормалев Д. А., Куршев Е. П., Сулейманова Е. А., Трофимов И. В.* Извлечение информации из текста в системе ИСИДА-Т // Труды 11-й Всероссийской научной конференции RCDL'2009. Петрозаводск: 2009. С. 247–253.
18. *Алексеев С. С., Морозов В. В., Симаков К. В.* Методы машинного обучения в задачах извлечения информации из текстов по эталону // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды XI-ой всероссийской научной конференции (RCDL'2009). Петрозаводск: КарНЦ РАН, 2009. С. 237–246
19. *Bird S., Klein E., Loper E.* Natural Language Processing with Python. USA: O'Reilly, 2009.

Быстрословарь: предсказание морфологии русских слов с использованием больших лингвистических ресурсов

Bystroslovar': morphological prediction new Russian words using very large corpora

Сокирко А. В. (sokirko@yandex-team.ru)

ООО «Яндекс», Москва, Россия

Рассматривается система предсказания морфологических признаков и полной парадигмы слова, разработанная для поиска по веб-страницам на русском языке. Система использует 12 факторов предсказания, размеченный корпус примеров, на котором тренируется модель машинного обучения. Описаны факторы, на которых строится обучение, делается попытка оценить значимость каждого фактора для решения поставленной задачи. Приводится сравнение построенной модели по точности и полноте с моделью, построенной на одном факторе.

1. Постановка задачи

Роль крупных Интернет-порталов, предоставляющих пользователям возможность поиска по вебу, растет с каждым годом. Ранжирование веб-страниц по релевантности пользовательскому запросу не решается без установления того, что слова из пользовательского запроса равны по значению словам, взятым с веб-страницы. Самое частое преобразование слов при поиске в рунете — морфологическое. Традиционно большая часть морфологической информации записывается в словаре, это удобно, поскольку морфология языка достаточно статична. Для каждого входного слова в этом словаре даются морфологические характеристики и парадигма словоизменения. Формально для поиска нужна только парадигма словоизменения, но в реальности определение, например, части речи может помочь установлению всех форм слова. Кроме статичного основного словаря, во многих системах строится специальный модуль предсказания новых слов, которых нет в основном словаре. Поскольку веб-поиск в каждый момент времени имеет дело с фиксированным («замороженным») корпусом, представляется целесообразным спроектировать словарь, функционально равный основному морфологическому словарю, но построенный автоматически по пользовательским запросам и веб-страницам. Такой словарь мы будем называть «быстрословарем», поскольку строится быстрее ручного словаря и поскольку он должен обновляться достаточно часто. В идеале этот словарь должен включать все слова, найденные в рунете и не вошедшие в основной словарь.

Данная публикация описывает новый алгоритм построения быстрословаря. Основным достоинством этой версии мы считаем:

1. Учет многих факторов морфологического предсказания внутри одной системы машинного обучения, что для русского языка не является разработанным приемом.
2. Получение информации из агрегированных пользовательских запросов и из текстов веб-страниц.

2. Обзор литературы

Системы морфологического анализа — одна из самых разработанных областей компьютерной лингвистики. Морфологические системы для русского языка традиционно основываются на грамматическом словаре А. А. Зализняка[1]. Алгоритмы предсказания обобщают морфологические схемы, предложенные в этом или родственном ему словарях, и/или учитывают распространенность этих схем в словаре и в обучающем корпусе. Один из первых алгоритмов морфологического предсказания для русского языка был предложен в работах Г. Г. Белоногова [2, 3]. Главным в этом алгоритме предсказания был принцип «корреляции между грамматическими признаками слов и буквенным составом их кон-

цов». Дальнейшее развитие алгоритмов, описанное, в частности, в работах Гельбуха[4], включало более детальную проработку правил композиции морфем и учет статистики морфем в корпусе текстов.

В 80-е и 90-е годы на факультете ВМК МГУ активно разрабатывалась система TULIPS-2, которая включала морфологический компонент[5], эта система использовала для предсказания словарь основ и словарь флексий, учитывались чередования.

Развитие корпусной лингвистики подстегнуло рост интереса к системам, которые в качестве решающего фактора используют частотность тех или иных морфологических схем в текстовом корпусе. Например, в работе Wicentowski[8] исследуется система, которая построена на трех простых факторах:

1. Расстояние Левенштейна, модифицированное под поиск морфологических вариантов
2. Контекстная близость по соседним словам в корпусе
3. Близость по частоте форм в одинаковых моделях словоизменения. Показывается, что система дает точность лемматизации порядка 80 % на 30 различных языках.

В работе Ножова [9] предлагается взвешивание гипотез морфологического предсказания с использованием метод корреляции. Матрицы корреляции строятся для основ и значений классифицирующих грамматических категорий. Гипотезы, имеющие максимальную корреляцию объявляются наиболее правдоподобными.

В работе Ляшевской и др.[10] был предложен метод взвешивания морфологического предсказания, основанный на следующем утверждении.

Если некоторое слово открытого (словоизменительного) класса встретилось в тексте в форме X, то скорее всего оно встретится в тексте в форме Y, отличной от первой. Из этого можно сделать предположение, что парадигмы новых слов тем лучше, чем больше разных форм этой парадигмы найдено в корпусе. В этой работе строились парадигмы для слов из Национального корпуса русского языка (НКРЯ).

В последних работах Goldsmith[11], одного из самых увлеченных исследователей в этой области, исследуется принцип минимального описания, который гласит, что морфологическая теория, которая описывает тот или иной корпус, должна быть минимальной по длине. Например, если в корпусе можно выделить N основ и K флексий, тогда такая теория лучше, чем теория, которая выделяет N+1 основ или K+1 флексий.

Хочется отметить, что многие работы в этой области затрагивают два аспекта, которые нас совершенно не интересуют:

1. Теоретическая морфология, когда авторы пытаются смоделировать очень глубокие законы, которые на сегодняшний день уже не являются продуктивными или распространенными. Мы считаем, если есть возможность задать «исключения» (следствие «устаревших» законов) списком, это достаточно.
2. Проблемы алгоритмической эффективности определения новых слов (нас интересует исключительно качество найденного, но не скорость работы программы).

3. Общая схема предсказания

Итак, главная задача — построить автоматический словарь («быстрословарь») по образу основного словаря. Основной словарь — это словарь Mystem[12]. Для нашей задачи можно считать, что это просто модифицированная версия грамматического словаря Зализняка. Каждое слово описывается набором форм и морфологическими характеристиками, которые приписаны этой форме. Например, для слова «мама»:

мама S,од,жен,ед, им,
мамы S, од, жен,ед, род
маме S,од,жен,ед, дат
маму S, од,жен,ед,вин
мамой S, од, жен, ед,твор
маме S,од,жен,ед,пр
мамы S,од,жен, мн,им
....

В каждой такой парадигме можно выделить псевдооснову (неизменяемую левую часть), в данном случае *мам-*, можно выделить StemGrammar (словообразовательные пометы, в данном случае «S,од,жен») и FlexGrammar (словоизменительные пометы). Можно записать данную парадигму в виде тройки <Основа, StemGrammar, Модель окончаний>, где модель окончаний — это набор пар вида <окончание, FlexGrammar>, например:

мама = <*мам*, «S,од, жен», F>,
где F = <-*а*, ед,им>, <-*ы* ед,род>..

В текущей версии словаря используются около 3000 моделей. Некоторые из них уникальны, например, есть специальная модель для слова *Комсомольск-на-амуре*, там выделяются окончания *-а-на-амуре*, *-ом-на-амуре* и т. д.

Предполагается, что в быстрословаре не будет отличных от основного словаря наборов окончаний или StemGrammar, предполагается даже, что самые редкие модели окончаний или StemGrammar будут убраны из рассмотрения.

Кроме основного словаря, используется еще два источника. Это веб-страницы рунета (около 4 миллиардов русских страниц без спама, только html) и лог частотных пользовательских запросов, где для

каждого запроса указана частота за месяц. Топ этого списка (всего 400 млн. запросов) выглядит так:

одноклассники	3 308 624
в контакте	1 457 124
порно	1 129 705
mail.ru	1 063 249
вконтакте	690 068
контакт	558 708
погода	458 272
зайцев нет	441 107
однокласники	391 875
работа	356 324
vkontakte	350 099
из рук	*в руки* 325 295
гороскоп	309 147
википедия	307 352
рамблер	303 245

По вебу и по запросам составлены частотные словари, где указана уже статистика для отдельных слов.

Кроме входных данных, есть еще корпус размеченных запросов (т. н. «морфотест»), в котором для каждого слова запроса указана вся парадигма слова. Корпус содержит 10 тысяч запросов и размечен вручную.

Общая схема построения словаря выглядит так. Мы берем все словарные слова из морфотеста, кроме стоп-слов и самых частых слова. Для каждого слова строятся все возможные морфологические интерпретации, хотя известно, что только одна из них верная. Для всех интерпретаций строятся значения всех факторов предсказания, и все вместе это составляет обучающую выборку.

Например, для слова *мама* строятся варианты (гипотезы):

Example	ModelId	StemGrammar	FlexLen	ModelFreq	IsGood
рома	7	S,имя,муж,од	1	212	–
Дюма	1135	S,од,фам	0	234	–
Фатима	7	S,жен,имя,од	1	2754	–
Айова	23	S,муж,неод,гео	1	13 908	–
ведьма	7	S,жен,од	1	1000	+
…					

Столбец Example — это пример, по которому была построена данная гипотеза. ModelId — номер модели окончаний (по нему строятся все формы слова). FlexLen — длина псевдоокончания, ModelFreq — частота модели окончаний с данным StemGrammar в основном словаре. Столбцы FlexLen и ModelFreq даны как примеры факторов предсказания, по которым происходит обучение (подробнее ниже). Столбец IsGood содержит целевой фактор обучения, он равен '+' только для верной гипотезы («S, жен, неод», ModelId=7). На самом деле в реальной модели целевой фактор является не булевским, а некой шкалой, поскольку нас интересует не точное попадание в единственно верную модель окончаний, а хотя бы приблизительное. В обучающую выборку было взято 4972 слова, по ним было построено 124 677 гипотез. Тестовая выборка состояла уже из несловарных слов морфотеста (998 слов), по ним было построено 24 923 гипотезы. Здесь может возникнуть вопрос: почему обучающая выборка состоит из словарных слов, а тестовая из несловарных? Очевидно, что выборки различны (см. об этом более подробно в работе Клышинского[13]). К сожалению, нам приходится мириться с этим, поскольку выборка по несловарным словам слишком мала для выделения из нее обучающей части.

Дальше на обучающей выборке строится модель машинного обучения, которая может предсказать StemGrammar и ModelId. На тестовой выборке проверяется качество модели. После этого мы берем топ самых частых несловарных слов из запросов (около 1,5 млн), строим для них все гипотезы, модель машинного обучения предсказывает StemGrammar и ModelId, по самым лучшим предсказаниям строятся парадигмы слов, если парадигмы сильно пересекаются друг с другом или со словарными словами, происходит отсев худших гипотез. Полученное множество парадигм и есть новый быстрословарь.

4. Факторы предсказания

Отдельный фактор должен голосовать за ту или иную гипотезу морфологического предсказания. Гипотеза морфологического предсказания (как сказано выше) — это модель окончаний (ModelId) и словообразующие характеристики (StemGrammar). Нет необходимости понимать, как отдельный фактор участвует в формуле предсказания, достаточно знания, что значение этого фактора **влияет** на выбор морфологической интерпретации. Оценка влияния фактора (importance, важность) возникает позже, после автоматического подбора формулы. Конечно, оценка Importance может только приблизительно показать значимость фактора, и она очень сильно зависит от выбранной модели машинного обучения. В нашем случае Importance дается по модели RandomForest[14], построенной на 30 решающих деревьях (реализация пакета R[15]).

5. Факторы внутреннего состава слова

Исторически самые используемые факторы для русского языка были связаны с окончанием сло-

ва (=псевдоокончание или постфикс). В нашей работы рассматриваются постфиксы длиной от 1 до 6 символов. По входному слову и гипотезе предсказания строится, соответственно, не больше 6 постфиксов, каждый из которых оценивается с помощью следующих параметров: Fp(p) — сколько раз постфикс p был использован для данной модели окончаний во всем словаре, Fg(p) — сколько раз постфикс p был использован с данным StemGrammar во всем словаре, Fa(p) — сколько раз постфикс p был использован во всех словах словаря. С помощью этих частот строятся следующие факторы:

Suffix_g — сумма всех Fg(p)/Fa(p) для всех постфиксов слов предсказываемой парадигмы.
BadSuffix_g — количество всех постфиксов, для которых Fg(p)=0 для всей предсказываемой парадигмы.

Suffix_p, **BadSuffix_p** — аналогично Suffix_g и BadSuffix_p, но вместо функции Fg используется Fp. Получается, что **Suffix_g** и **BadSuffix_g** оценивают, насколько слово подходит предсказываемому StemGrammar, а **Suffix_p**, **BadSuffix_p —** предсказываемой модели окончаний.

Важность этих факторов для предсказания StemGrammar высока:

Фактор	Предсказание StemGrammar	Предсказание ModelId
Suffix_g	20,21 %	7,09 %
Suffix_p	2,54 %	6,40 %
BadSuffix_g	8,07 %	2,84 %
BadSuffix_p	3,12 %	3,11 %

Суммарная важность этих факторов равна 33 % для предсказания StemGrammar и 20 % для предсказания ModelId.

Кроме этого, еще использовались другие факторы, описывающие внутренний состав слова:

- **FlexLen** — длина псевдоокончания для предсказываемой формы внутри предсказываемой парадигмы.
- **Hasprefix** — в словаре найдено другое слово W той же модели окончания, такое что входное слово делится на две части P и W, где P — продуктивный префикс типа «квази», «экс» и т. д.
- **lemma_leven** — в словаре найдено другое слово W той же модели окончания такое, что расстояние Левенштейна между входным словом и словом W не больше некоторого порога.
- **AbbrLike** — отношение числа согласных букв слова к общему количеству букв слова.

Каждый из этих пяти факторов имеет важность не больше 2 %. На данном этапе не очень понятно, почему их вклад такой незначительный.

6. Контекстные факторы

Следующие два фактора используют данные про контексты слов, взятые с web-страниц. Под обработку контекстов была специально адаптирована модель Trigram[16], построенная на полных тегах и обученная на Национальном корпусе русского языка. Для каждого контекста входного несловарного слова строятся вероятности того, что этому слову может быть приписана данная морфологическая интерпретация (StemGrammar и FlexGrammar). При использовании этих факторов возможны следующие осложнения. Во-первых, тексты НКРЯ не равны текстам рунета по жанрам и среднему количеству ошибок, хотя это не так критично, ведь модель Trigram основывается на тройках морфологических интерпретаций, типа

«А,ед,муж,им»,
«А,ед,муж,им»,
«S,ед,муж,им» *// красивый красный шар*
// новый желтый стол

Такие триграммы широко распространены как в НКРЯ, так и в рунете.

Во-вторых, некоторые контексты в рунете столь неоднозначны, что опираться на них не стоит. Учитывая это, мы пропускаем те контексты, в которых уровень морфологической неоднозначности превышает некий порог. В-третьих, не все тексты одинаково хороши, некоторые сайты содержат много мусора. Оценить значимость сайта можно с помощью т. н. тематического индекса цитирования[17], но пока этого не сделано.

В текущую модель включено два контекстных фактора:

- **weight_context_exact** — количество контекстов рунета, которые проголосовали за морфологическую интерпретацию входного слова;
- **weight_context_sum** — считаем **weight_context_exact** для каждой формы парадигмы и нормируем на число форм в парадигме.

Суммарная важность этих двух факторов находится в районе 12 %.

7. Факторы частот

Следующая группа факторов связана с частотными словарями рунета и пользовательских запросов:

- **QF** — частота входного слова в пользовательских запросах;
- **TF** — частота входного слова в рунете;
- **PROP** — отношение числа раз, когда слово было записано в рунете с большой буквы, к значению TF;

- **QF1** — частота слова в однословных запросах;
- **PrdFreqModel** — среднее квадратичное отклонение относительных частот форм этой модели окончания. Например, в парадигме слова «мама» (ModelId=7), относительная частота словоформы *мама* — 30 % , а словоформы *маме* — 10 % и т. д. Если в этой модели (ModelId=7) средние относительные частоты совсем другие — тогда штрафуем эту гипотезу.
- **paradigm_found_pure** — число форм парадигмы, найденных в рунете, поделенное на число форм парадигмы.
- **lemma_query_freq** — число вхождений предсказываемой леммы парадигмы в лог пользовательских запросов.

Суммарная важность этих факторов для StemGrammar — 25 %, а для ModelId — 35 %. Именно по этим группам факторов происходит основное расхождение между словарными и несловарными словами, поскольку, например, словарные в среднем используются в рунете на порядок чаще, чем несловарные.

8. Остальные факторы

Поскольку среди морфологических помет есть несколько, которые не относятся непосредственно к морфологическим (гео, имя, фам), нам пришлось подключить два источника:

1. Тезаурус географических названий Яндекс.Карт.
2. Частотные словари сервиса МойКруг для фамилий и имен.

На основе этих перечней мы построили фактор **SemFeat**, который, например, поднимает гипотезы с пометой «фам», если это слово часто встречается на сервисе МойКруг. Важность этого фактора не оказалась высокой даже для предсказания StemGrammar (1 %).

Кроме этого, мы использовали еще два дополнительных технических фактора:

- **ModelFreq** — частота модели окончаний в основном словаре (важность — 2 %).
- **FormsCount** — число форм в предсказываемой парадигме (важность — 20 %).

9. Анализ результатов

Построение всех факторов занимает порядка 20 часов на кластере в 200 компьютеров (каждый компьютер — 8 процессоров, 20 ГБ ОЗУ, 5 ТБ диск). Основное время уходит на подсчет контекстных факторов.

Сравнение по качеству производилось по двум метрикам:

1. Точность определения StemGrammar;
2. Точность и полнота определения форм парадигмы.

Последние определялись как средние точность и полнота для каждого слова запроса:

$$Точность: P = C/V,$$

$$Полнота: R = C/Q,$$

где C — сумма частот(в рунете) правильно предсказанных форм парадигмы, V — сумма частот всех предсказанных форм парадигмы, Q — общая сумма частот всех правильных форм парадигмы.

$$F\text{-}Measure: Fm = 2*P*R/ (P+ R)$$

Для сравнения была использована предыдущая версия быстрословаря, которая по большей части базировалась на работе[10]. Ниже приведены показатели качества:

	Точность	Полнота	F-measure
StemGrammar, old	0,6716		
StemGrammar, new	0,7517		
Формы, old	0,9295	0,9372	0,9354
Формы, new	0,9282	0,96	0,9439

Прирост точности по StemGrammar достаточно большой (с 67 % до 75 %), это вполне соответствует нашим ожиданиям, ведь мы перешли от довольно простой модели (которая включена в нашу модель в виде одного из фактора частот) к многофакторной модели, учитывающей многие аспекты использования слова.

Показатели по формам не столь оптимистичны. Общая F-measure выросла на процент, однако точность немного снизилась (это не мешает нам говорить о том, что модель в целом лучше предыдущей). Почему точность снизилась? Ниже приведен список самых грубых ошибок нового быстрословаря на текущей версии морфотеста:

Дробо	Предсказано как краткое прилагательное, должно быть фамилией
пожаро	Должно быть в основном словаре
стил	Должно быть неизм. фам, а стало изм., мешают формы *стиле*
руссо	Стало изменяемым, формы мешаются с фамилией *Русс*
Мега	Иногда изм, иногда нет
Дони	предсказалось как имя *Доня*, а в запросе это часть детской считалки
Волошко	не стало фамилией

Куинджи	Предсказано как изменяемая фамилия
Макроуровне	предсказалось как отчество по *Артуровне*
Рогово	Предсказалось как прилагательное
Судоку	Предсказалось как изменяемое

Даже этот короткий список примеров может определить направления будущей работы:

1. Проблемы с «Пожаро». Основной словарь требует доработки, не расширения, а упорядочивания. Например, мы последовательно должны включать композитные формы в парадигмы прилагательных.
2. Многие формы в интернете (судоку, куинджи) становятся изменяемыми. Например, таковым стало слово «вконтакте» (вконтакту, «вконтактом»…). Провести границу между совсем уже редким или полушуточным использованием или привычным освоением этого слова трудно, но нужно.
3. Пересечение слов со словарными и предсказанными парадигмами (*стил*, *руссо*), особенно для коротких слов, требует отдельной проработки. Фактически почти все предложенные факторы — это факторы для жадного захвата форм в парадигмы, но должны существовать факторы, которые препятствуют объединению несовместимых форм. Единственный фактор, который используется для этого, — это PrdFreqModel. Возможно, нужно вводить еще факторы, например, сколько раз разные формы слов одной парадигмы встречались на одной веб-странице.
4. По всей видимости, любые контекстные факторы требуют домножения на оценку качества самой веб-страницы.

Кроме тактических соображений, еще предстоит начать переоценку масштабов заимствования и скорости освоения новых слов рунета.

Благодарности

Я выражаю благодарность всем сотрудникам отдела лингвистических технологий компании «Яндекс» за помощь в проведении этого исследования, а особенно:

1. Евгению Соловьеву (машинное обучение);
2. Николаю и Светлане Григорьевым (разработка системы оценки качества);
3. Елене Грунтовой (общее руководство и идея исследования);
4. Вере Цукановой (основной морфологический словарь);
5. Андрею Кондратьеву (предыдущая версия быстрословаря).

Литература

1. *Зализняк А. А.* Грамматический словарь русского языка. Словоизменение. — М.: Русский язык, 1977.
2. *Белоногов Г. Г.* Об использовании метода аналогии при автоматической обработке текстовой информации // Проблемы кибернетики. — М.: 1974, вып. 28.
3. *Белоногов Г. Г., Зеленков Ю. Г.* Алгоритм морфологического анализа русских слов // Вопросы информационной теории и практики. № 53. Автоматическая словарная служба. Автоматическое индексирование документов. М., 1985. С. 62–93.
4. *Гельбух А. Ф.* Эффективно реализуемая модель морфологии флективного естественного языка: Автореф.дис… к. ф. н. / АНРФ., ВИНИТИ. — М.,1994.
5. *Мальковский М. Г., Волкова И. А.* Анализатор системы TULIPS-2. Морфологический уровень // Вестн. Моск. Ун-та, сер. 15, 1981, N 1, с. 70–76.
6. *Шереметьева С. О., Ниренбург С.* 1996 Эмпирическое моделирование вычислительной морфологии. //НТИ, №7, 1996.
7. *Goldsmith, J.* Unsupervised Learning of the Morphology of a Natural Language. //University of Chicago, 1998.
8. *Wicentowski, R.* Modeling and Learning Multilingual Inflectional Morphology in a Minimally Supervised Framework, 2002.
9. *Ножов И. М.* Реализация автоматической синтаксической сегментации русского предложения. Дисс… канд. тех. наук. — М.: РГГУ, 2003.
10. *Ляшевская О. Н., Сичинава Д. В., Кобрицов Б. П.* Автоматизация построения словаря на материале массива несловарных словоформ // Браславский П. И. (отв. ред.), Интернет-математика — 2007: сб. работ участников конкурса науч. проектов по информ. поиску. Екатеринбург: Изд-во Урал. ун-та. С. 118–125.
11. *Goldsmith, J.* 2001. The unsupervised learning of natural language morphology. Computational Linguistics 27(2): 153–198.
12. *Сегалович И., Маслов М.* Русский морфологический анализ и синтез с генерацией моделей словоизменения для не описанных в словаре слов // Труды международной семинара Диалог'98 по компьютерной лингвистике и ее приложениям. Казань, 1998. Т. 2. С. 547–552.
13. *Клышинский Э. С.* Некоторые сложности автоматизированной лемматизации несловарных словоформ [Текст] [Электронный ресурс] / Э. С. Клышинский // Материалы международной конференции «Диалог 2009». — М.: РГГУ, 2009
14. *Breiman, L.* Random forests. Machine Learning, 45(1): 5–32, 2001. 18
15. *Liaw, A., Wiener, M.* Classification and regression by randomForest. Rnews 2002, 2:18–22.
16. *Сокирко А. В., Толдова С. Ю.* Сравнение эффективности двух методик снятия лексической и морфологической неоднозначности для русского языка // Интернет-математика-2005.
17. *Что такое тИЦ?* Электронная публикация. http://help.yandex.ru/catalogue/?id=873431
18. *Toutanova, K., Cherry, C.* 2009:A global model for joint lemmatization and part-of-speech prediction

Корпусное исследование лексико-семантических отношений между 6 русскими словами, обозначающими капитальные объекты (контексты с однородностью)

Corpus-based evidence for lexico-semantic relations for 6 russian words designing permanent structures

Соколова Е. Г. (minegot@rambler.ru)

РГГУ, Москва

В статье описываются результаты корпусного исследования лексико-семантических отношений между шестью словами русского языка: дом, здание, строение, постройка, корпус и сооружение. Предлагается методика исследования, основанная на классификации контекстов, в которых эти слова совместно встречаются в корпусе текстов. Перечисляются синтаксические конструкции, представляющие контексты с однородными объектами, и ассоциируемые с ними лексико-семантические отношения.

1. Введение

Задача, решаемая в данной статье связана с исследованиями по автоматической генерации описаний изображений на ЕЯ (1, 2, 3, 4), в которых авторы системы столкнулись с проблемой выбора номинаций для упоминаемых в тексте домов и других капитальных объектов и их описаний в онтологии, используемой генератором при порождении текстов.

Систематизация понятий капитальных объектов для генерации описаний изображений обслуживает несколько процессов. Назовем три: 1) первичная номинация объектов, основанная на зрительно воспринимаемых свойствах объекта или зрительном образе, 2) упоминание в тексте понятия, отсылающего к уже введенному в дискурс понятию (вторичная номинация) и 3) наследование свойств от более общих к более конкретным понятиям в онтологии. Одно и то же слово может использоваться как первичная и как вторичная номинация. Например, **[1]**[1] *«Повернув за угол этого здания, любопытствующий посетитель или вновь прибывший арестант видит в дальних углах двора два одноэтажных строения, гораздо меньших, чем главный корпус.»* — первичная номинация объектов словами *строение* и *корпус*. В следующем контексте *дом* — первичная номинация, *строение* — вторичная: *«Ты назначишь ему свидание в квартире, которая расположена в этом доме. — Гречишников указал на сталинское строение, увенчанное стеклянным газовым ангелом*, где между словами *дом* и *строение* существует родовидовое отношение или отношение синонимии.

Исследуются связи пяти слов, описываемых в словарной статье ДОМ 1 Нового Объяснительного Словаря Русского Языка (НОССРЯ) (5, с. 82–84) как синонимы, каковыми они могут являться в определенных своих значениях, и которые удовлетворяют следующему смыслу[2]: «наземное сооружение, имеющее внутри помещения, которые занимают большую часть его объема», сформулированному для них в этой словарной статье. Подход НОССРЯ близок подходу WordNet (6), в котором похожие слова объ-

[1] В статье сквозная нумерация примеров. Нумеруются только примеры, которые далее рассматриваются в р.6. Примеры, иллюстрирующие не рассматриваемые и терминологические контексты, не нумеруются. Номер ставится перед примером в квадратных скобках. Длинные примеры сокращены без ущерба для рассматриваемого явления, удаленные куски заменены многоточием.

[2] В статье используется термин «смысл» в отличие от термина «толкование», используемого в словарях типа НОСС (см. ниже ссылку 4), поскольку обсуждается именно употребление слов в текстах.

единяются в «синсеты» по общему значению/смыслу. В WordNet найдено удачное решение — созданы синсеты и формализованы лексико-семантические отношения (ЛСО) между ними, что превращает традиционный словарь в электронный ресурс. Прочая информация, в частности, толкования, остаются традиционными. В отличие от WordNet, представляющего собой интернет-ресурс и полную лексико-семантическую базу английского языка, которая свободно доступна в Интернете для использования людьми и автоматизированными системами, НОССРЯ имеет вид традиционного филологического исследования, адресованного специалистам, и представляет собой фактически сборник исследований, более или менее субъективных в зависимости от слова (служебное / полнозначное) и автора статьи. Наша попытка опереться на НОССРЯ для включения в онтологию понятий капитальных объектов показывает, что статья ДОМ 1 не является достаточной базой для моделирования употребления слов в тексте. Традиционный словарь дает мало информации, выделяя только точные синонимы. Например, в (7) две пары из десяти комбинаторно возможных для вышеперечисленных пяти слов, считаются синонимами — *строение* и *постройка*, *здание* и *дом*.

Цель исследования состоит в том, чтобы найти объективные основания для различения ЛСО между двумя словами с позиции их употребления в тексте. Материалом исследования являются контексты, найденные в Национальном Корпусе Русского Языка (НКРЯ) (8) для пар слов из рассматриваемой группы, состоящей из пяти вышеперечисленных слов и слова *сооружение*, которое по его месту в определении смысла синсета, данного в НОССРЯ, является гиперонимом этих слов, следовательно, его смысл пересекается со смыслом данного синсета. По нескольким поисковым запросам из НКРЯ извлекаются и исследуются контексты совместного употребления пар слов. Возможны два типа контекстов: а) двумя словами обозначаются разные, но однородные (в силу принадлежности к одной семантической группе) объекты — контекст однородности; б) двумя словами обозначается один и тот же объект — контекст тождества. Размер статьи не позволяет рассмотреть оба типа контекстов, поэтому в данной статье мы ограничиваемся контекстом однородности. Кроме того, не рассматриваются контексты, близкие к тождественным, в которых понятия, обозначенные парой слов, находятся в различных ракурсах (см. р. 3).

2. Метод (часть 1)

Рассматриваются все возможные кортежи пар слов из заданного набора, всего 30 кортежей. Для определения наличия между двумя словами родовидового отношения и синонимии обычно используются психолингвистические тесты: а) тест на заменимость слова в тексте его синонимом — для синонимии, б) диагностические контексты: «*Это X, следовательно, это Y*» и «*Это Y, следовательно это X*» для обоих отношений. Если верны оба утверждения, то диагностируется отношение синонимии, если одно верно, а другое нет — родовидовое отношение, например, слово *птица* — гипероним слова *соловей,* потому что (верное утверждение отмечается «t», неверное — «f»): t(*Это соловей, следовательно, это птица*) и f(*Это птица, следовательно, это соловей*). *Судно* — синоним *корабль*, потому что: t(*Это корабль, следовательно, это судно*) и t(*Это судно, следовательно, это корабль*). Однако опереться на диагностический контекст в нашем случае не удается по двум причинам: а) значения трех из шести слов ассоциируются с объектом (и с некоторым зрительным образом[3]): *дом, здание, корпус*, а другие воспринимаются как результат созидательного процесса, т. е. ассоциируются с процессом, например, «сооружение — это «то, что сооружено» (9). Для слов с объектной ассоциацией слова с процессной ассоциацией окажутся гиперонимами: t(*Это дом / здание / корпус, следовательно, это строение / постройка / сооружение*) по известному нам способу возникновения объектов — они *построены, сооружены*. Но знание не тождественно ЛСО языка. Гипотеза, проверяемая в данной статье, состоит в том, что ЛСО должны проявляться в текстах. Диагностические контексты в реальных текстах не встречаются, возможности проверить взаимозаменяемость слов нет. В статье рассматриваются контексты, в которых исследуемые слова встречаются совместно, найденные в НКРЯ по трем следующим шаблонам:

[X (1 1) Y] — выделяет контексты, в которых слова рассматриваемой пары следуют непосредственно друг за другом (знаки препинания в НКРЯ не считаются словом), например, [дом (1 1) здание], по которому найдется, в частности, контекст: **[2]***Рабóта архитéкторов видна всем: вот онú, их домá, здáния.* [Михаил Песин. Картинки с выставки (2002) // «Биржа плюс свой дом» (Н. Новгород), 2002.05.20].

[X (1 1) это (1–6) Y] и [X (1 1) быть (1–6) Y] — шаблоны, предназначенные для выделения конструкции типа «определение», например, «сосна — это дерево». В связи с тем, что между словом «это/быть» и вторым словом пары допускается вставка до 6 слов, эти шаблоны выделяют дополнительные контексты — с указательным местоимением *это*, с глаголом *быть* в прошедшем времени, и др.

В табл.1 приведены данные о количестве найденных в НКРЯ контекстов для всех пар слов. Ко-

[3] Можно сказать: «*Нарисуй дом, здание*», плохо «*нарисуй корпус*», и совсем невозможно «*нарисуй сооружение, строение, постройку*».

лонка «?» содержит сведения о количестве не интерпретируемых однозначно примеров. Колонка «Не рассм.» содержит контексты, в которых слова имеют значения, отличные от смысла синсета. Это примеры со словами *строение, дом* и *корпус* как элементами адреса; слова *строение, постройка* и *сооружение* как процесс, например, *сооружение домов, корпус постройки 19 в.*; *корпус* как оболочка, например, *И тут Христо заме́тил в темноте́ — на мину́ту совсе́м я́сно — огро́мные, как облака́, паруса́ и высо́кий, как дом, ко́рпус*. [Б. С. Житков. Элчан-Кайя (1926)], *сооружение* как устройство.

Табл. 1

		Всего	?	Не рассм.
1	дом здание	14	1	
2	дом строение	4	1	
3	дом постройка	11		10 *дома постройки 70ых годов*
4	дом корпус	11		9 — адрес (см.)
5	дом сооружение	1		
6	здание дом	39	2	30 *здание дома престарелых*
7	здание строение	84		
8	здание постройка	15	1	13 *здание постройки прошлого века*
9	здание корпус	9		
10	здание сооружение	107		
11	строение дом	22		12 (устар.) *строение домов*
12	строение здание	6		
13	строение постройка	-		
14	строение корпус	1		
15	строение сооружение	65		
16	постройка дом	101		94 *постройка дома, старой постройки дом,…*
17	постройка здание	19		16 *постройка зданий, старой постройки здание*
18	постройка строение	-		
19	постройка корпус	-		
20	постройка сооружение	2		
21	корпус дом	19		16 *корпус дома*
22	корпус здание	7		1 *корпус здания*
23	корпус строение	2		— (устар.)
24	корпус постройка	1		1 *корпус постройки 18 века*
25	корпус сооружение	-		
26	сооружение дом	11		8 *сооружение дома*
27	сооружение здание	3		2
28	сооружение строение	3		1
29	сооружение постройка	3		
30	сооружение корпус	-		

Не рассматриваются контекты, в которых слово сохраняет ассоциацию с процессом, например, а) *В то же время Петербург стал украшаться и новыми, наконец достойными и гармонирующими с характером Петербурга зданиями — постройками молодых архитекторов Фомина, Щуко, Таманова и др.* [М. В. Добужинский. Облик Петербурга (1938) // «Звезда», 2003]. б) *К городу примыкали громадные постройки домов и судостроительных верфей*. [А. С. Новиков-Прибой. Цусима (1932–1935)]. Пример а) можно перефразировать: «*зданиями постройки молодых архитекторов*», где слово *постройка* будет иметь только процессное значение, в б) реализуется устаревшее значение слова *постройка* — «место, где строят» (9).

Пустые строки табл. 1 интерпретируются в р. 4. В строках 7, 10 и 15 особенно много примеров. Это административные штампы. Они рассматриваются в р. 5. Оставшиеся контексты интерпретируются по типам и ЛСО для каждой пары слов в р. 6.

3. Метод (часть 2)

Рассматриваются попарно вышеперечисленные слова, и проверяется наличие родовидового отношения и синонимии между членами этих пар. Обозначим отношения стрелками:

X → Y- родовидовое отношение: Y — гипоним, X — гипероним;
X ↔ Y- синонимия: Y и X — синонимы.

Пробуем найти манифестацию этих отношений в контекстах из НКРЯ. Заметим, что слова выделены на основании смысла синсета и точными синонимами не являются. Наоборот, в НОССРЯ автор словарной статьи ДОМ 1 пытается «развести» их значения, указывая различительные признаки, которыми обычно обладают обозначаемые ими объекты: размер, материал и др[4]. Различительные признаки рассматриваются как часть лексического значения слова. Они сообщают гипониму конкретность по отношению к гипероним в случае, если между смыслами слов обнаруживается родовидовое отношение, например, *дом* имеет некоторые конкретные свойства, которые выделяют его как вид *наземных сооружений* и отличают от других видов наземных сооружений *(здание, строение* и др.), вступающих с ним в отношение квазисинонии. Получается, что оба отношения, родовидовое и квазисинонимии, базируются на понятии различительного признака. Отношение квазисинонимии обозначим разорванной стрелкой: X←{_|_}→Y, где в фигурных скобках помещаются различительные признаки слов X и Y, соответственно. Если одновременно есть свидетельство о квазисинонимии и наличии родовидового отношения, будем обозначать эту ситуацию такой же формулой, но с удвоенной стрелкой, указывающей на гипоним: X←{_|_}= >>Y. Родовидовое отношение тоже может указывать различительный признак гипонима, например, «←+жилой|», или гиперонима, например, «←|+рез.процесса».

В данном исследовании оказалось полезным использовать комплексный признак, привязанный как к коммуникативной ситуации, так и к обозначаемой реальности, назовем его ракурсом. Приблизительно, ракурсы составляют две группы: застройка — когда речь идет об объектах вообще, объектах в регулирующих документах, объектах населенных пунктов определенного типа, и визуальный ракурс — когда объект представляется (как бы) в зрительном восприятии говорящего:

1 «застройка» — совокупность капитальных объектов, составляющих населенный пункт, например, *московские здания,* включая объект как элемента застройки, например, *здание, строение* и *сооружение* в документах, в частности, этот ракурс свойственен официально-деловой сфере функционирования текстов.

2 «мини-жил» — совокупность капитальных объектов в «распределенном капитальном объекте» (РКО), предназначенном для постоянного (в том числе сезонного) проживания — *усадьбе, дворе, даче*. Под РКО понимается объект, занимающий относительно небольшой и часто огороженный участок земли, объединенный общим собственником или пользователем, состоящий обычно из нескольких капитальных объектов;

3 «мини-спец» — совокупность капитальных объектов в РКО специального назначения — *монастыре, больнице, заводе* и др. Ракурсы **2** и **3** называем мини-ракурсами.

4 ракурс зрительного восприятия («ЗВ»), ракурс наблюдателя — застройка, не огороженная, например, застроенная *поляна, берег, совхозные угодья, дачный поселок*, зрительно воспринимаемые рассказчиком;

В данной статье рассматриваются контексты, в которых оба слова находятся в одном и том же ракурсе. Кроме того, что каждое слово пары обладает определенным ракурсом, оба слова погружены в определенный синтаксический контекст.

В данной статье мы рассматриваем контексты однородности, в которых соучаствуют слова одного синсета. Он создается при перечислении объектов в составе сочиненной цепочки или при наличии в предложении указателей, создающих единый ситуационный контекст для обоих слов, например, общее обстоятельство для двух сочиненных предложений, соединение сравнительным предикатом и др. Выделяется три типа контекстов однородности: O_сочинение, O_сочинение_с_обобщением и O_ — остальные контексты однородности.

O_сочинение — цепочка однородных членов, в которой однородность связана либо с принадлежностью членов одному синсету, либо наличием у них общего гиперонима. В первом случае однородные члены являются квазисинонимами с различительными признаками, например, [дома, здания] ➔ дом←{+жилой|-жилой}→здание; во втором имеют общего гиперонима, например, в терминосистеме «недвижимость» (см. ниже). Результат можно записать:

[X, Y] и (X и Y принадлежат одному синсету) ➔X←{a|b}→Y, где a и b — различительные признаки.

[4] Можно понять, что толкование и различительные признаки — это элементы метаязыка описания слова в словаре типа НОСС (Иомдин Б. Л. Как составляются словари. Конспект семинаров // Лингвистика для всех. Летние лингвистические школы 2005 и 2006. М: «МЦНМО», 2008. С. 85). Встречается два типа различительных признаков: маркированность, например, «+ — наличие архитектурного замысла» в статье ДОМ 1 НОССРЯ, и положение на шкале, по которой противопоставляются слова — члены синонимического ряда, например, для слов *тоска*, *уныние* и *грусть* по шкале «интенсивность, глубина и длительность чувства в вышеуказанном источнике с. 85. Мы тоже используем термин «различительный признак», причем в нашем случае это — маркированность. В статье «Синонимия» энциклопедии Кругосвет (без подписи) подобный термин не используется, просто речь идет о смысловых, стилистических и др. различиях http://www.krugosvet.ru/enc/gumanitarnye_nauki/lingvistika/SINONIMIYA.html.

[X, Y] и ¬(X и Y принадлежат одному синсету) ➔найдется Z (Z →X и Z →Y).

O_сочинение_с_обобщением — сочиненная цепочка, вводимая обобщающим словом, причем обобщающее слово и некоторые члены цепочки принадлежат одному синсету, например, *«возведение каких-либо строений (домов, поселков, предприятий и т. д.)»*. Члены цепочки отделяются от обобщающего слова двоеточием, заключаются в скобки, иногда с уточнением, например, *в том числе*:

[Z (X, Y)] и (X и Z принадлежат синсету) ➔ Z→ X.

Разновидностью этого типа являются сложносочиненные предложения с квантором всеобщности в первом предложении

[S1(*все* Z), S2(X)] и (X и Z принадлежат одному синсету) ➔ Z→X, например, **[32].**

O_ — слова одного синсета обозначают однородных участников одной ситуации. Тип отношения между ними зависит от контекста. Это может быть синонимия: употребление для номинации слов–синонимов имеющее стилистическую мотивацию — исключение повторения слова:

[3] *Один из них, 50-летний безработный, позвонил в милицию из квартиры на Сумской улице и сообщил, что соседнее с его домом здание ОВД «Чертаново-Северное» заминировано.* [Дмитрий Павлов. Ведомости (1996) // «КоммерсантъDaily», 1996.01.20].

Оба объекта — дом безработного и здание ОВД, могут быть названы и одним, и другим словом, например, *«один из домов (одно из зданий) на Сумской улице заминирован(/о)»*[5]. Это может быть отношение часть, см. р. 6.4 *дом — корпус*, где содержание отношения создается определениями слова *корпус*. Можно указать на следующие синтаксические типы контекстов:

[Z Pr …X Y], где (Z — согласованное определение к Y), (Pr зависит от Z),

(X и Y принадлежат синсету) ➔ X ↔ Y, например, **[3]**.

[Z *как* …X Y], где (Z — согласованное определение к Y),

(X и Y принадлежат синсету) ➔ X ↔ Y:

[4] *Давно не крашенный, серый, как большинство старых деревянных строений, дом тем не менее носил следы хорошей хозяйской руки.* [И. А. Ефремов. Лезвие бритвы (1959–1963)]

4. Пустые строки табл. 1 (13 и 18, 19 и 24, 25 и 30)

Из табл. 1 видно, что пары слов «*сооружение* и *корпус*», «*постройка* и *корпус*», а также «*постройка* и *строение*» не встречаются в одном контексте.

Слова *строение* и *постройка* по тесту с диагностическим контекстом являются точными синонимами и включены в (7) и при этом не встречаются в одном контексте, который мог бы диагностировать родовидовое отношение или квазисинонимию. Следовательно, можно рассматривать отсутствие таких контекстов как подтверждение синонимии этих слов. Синонимии не противоречит факт, что в современном языке сферы употребления этих слов различны: *постройка* в отличие от *строение* крайне редко употребляется в текстах официально-деловой сферы.

Для пар «*корпус-постройка*», «*корпус-сооружение*» с разной природой значения диагностический контекст не показателен. В контекстах из НКРЯ для этих пар слов ЛСО не диагностируются, следовательно, их нет. Это утверждение не противоречит знанию того, что *корпус,* по сути, является *зданием* или *домом* (только называется *корпусом*) и как таковой ассоциативно может быть связан с понятиями, находящимися с ними в лексико-семантических отношениях. Учитывая ограниченность эксперимента также нельзя утверждать, что такие контексты в принципе невозможны, но в контекстах однородности эксперимента эта возможность не проявляется, и эти ассоциации не работают.

5. Здание, строение, сооружение (строки 7, 10, 15, 28) как термины

В строках 7, 10, и 15 табл. 1 оказалось особенно много примеров, причем подавляющее большинство встречаются в одном сочетании: «*здание, строение*, / и /или *сооружение*». Это сочетание является штампом официально-деловой сферы — документации по недвижимости. Строка 7 выделяет начальную пару слов из этой тройки (*здание строение*) — 84 контекста, строка 15 — конечную пару (*строение сооружение*) — 65 контекстов. *Дом* встречается в тройке в найденных контекстах вместо слова *здание* всего два раза, причем в «маргинальных» ситуациях — *заложены* или *самовольной постройки*. Кроме того, в этих двух примерах слово *строение* предваряется определением «другой, иной»: а) *Самовольной по-*

[5] Замена слов в данном примере невозможна, потому что *здание* не сочетается с притяжательным местоимением, обозначающим собственника.

стройкой является жилой дом, другое строение, сооружение или иное недвижимое имущество, созданное на земельном участке, не отведенном для этих целей …. [Гражданский кодекс Российской Федерации. Часть первая (1994)]; б) … *решением об изъятии земельного участка, на котором находятся заложенные дом, иные строения, сооружения или насаждения, убытки, …, возмещаются залогодержателю в полном* …. [Закон Российской Федерации «О залоге» (1992)]. Это свидетельствует о том, что в этих примерах реализуется отношение синонимии в паре «*дом–строение*» как терминов или «общеязыковой» синонимии слов *дом* и *здание* в несвойственной для *дома* сфере употребления в маргинальных для *здания* контекстах. Второе предпочтительнее.

Строка 10 содержит административный штамп «*здания, сооружения*», который используется часто в сочетании с «*земельным участком*» и представляет сжатый вариант трехчленного штампа — 107 контекстов; вариант двучлена со словом *строение* встречается только два раза. Термины встречаются только в одном типе контекстов — **O_сочинение.**

В табл. 2 представлены ЛСО официально-деловой сферы «недвижимость» для нашей группы слов.

Табл. 2

		застройка	
1	объект_недвижимостиТ	→	зданиеТ, строениеТ, сооружениеТ
2	зданиеТ, строениеТ	←{_|-синсет*, -процесс}→	сооружениеТ
3	зданиеТ	←{?|?}= >>	строениеТ

* *СооружениеТ* не принадлежит синсету и не имеет процессного значения, определенного для него в (9). Главная составляющая значения термина в том, что он представляет «объект недвижимости», т. е. является его гипонимом. *Дом* как объект недвижимости встречается крайне редко в основном для обозначения «неполноценного» объекта — самовольно построенного или заложенного.

Комментарии. В табл. 2 слова-термины помечены большой буквой «Т» в конце слова. Значения терминов сужены по сравнению с теми же словами — не терминами. *Сооружение* не имеет процессного значения, определенного для него в (9) — (-процесс). Главная составляющая его значения в том, что оно обозначает «объект недвижимости», т. е. является гипонимом этого понятия, и обозначает объекты инфраструктуры населенных пунктов, например, *очистные, спортивные сооружения, дороги и* т. п., отличные от *зданий и строений*. *Строение* является квазисинонимом термина *здание* с невыявленным различительным признаком и его же гипонимом.

6. Контексты с нетерминологическими значениями

6.1. Пара слов «дом и здание» (строки 1 и 6)

Обе части диагностического контекста для данной пары слов в целом истинны: (t)*Это дом, следовательно, это здание.* и (t)*Это здание, следовательно, это дом*. Т.е. слова могут считаться синонимами, что подтверждается и их наличием в (7).

O_ (6 контекстов) также свидетельствуют в пользу синонимии (в ракурсе dev, dev-el):

[5] *В городе много домов было разрушено, чувствовалась острая нехватка зданий.* [Любовь Кузнецова. «…Собираю разрозненные брёвнышки народа своего…» (2003) // «Вестник США», 2003.09.03]

[6] *Цены впрямую зависят от технического состояния объектов, и самые лучшие в этом смысле здания будут стоить не ниже сталинского дома.* [Квартиры б/у: не хуже новостроек (2003) // «Мир & Дом. City», 2003.09.15]

[7] *И только потому, что школьные дома были поместительнее других зданий, принадлежавших сельским коммунам, школы естественно служили местом собраний* …. [Ф. К. Сологуб. Королева Ортруда (1909)].

Пример в ракурсе «ЗВ»: **[4],** а также:

[8] *Они миновали уж здание; вон за зданием — дом; и вон — окна…* [Андрей Белый. Петербург (1913–1914)].

O_сочинение — наиболее частый (11 контекстов) для этой пары: см. **[3]**, а также:

[9] *Все его сооружения — жилые дома, здания городских учреждений, заводские цеха, проезжие и пешеходные дороги — покоились на сваях.* [Евгений Рубин. Пан или пропал. Жизнеописание (1999–2000)]

[10] *Я с ними рассматриваю рисунки древних зданий, домов, утвари, — сам черчу, объясняю, как, бывало, тебе: что сам знаю, всем делюсь.* [И. А. Гончаров. Обрыв (1869)]

[11] *Но декорация совершенно переменилась, когда мы выехали на берег Сены; тут представились нам красивые здания, домы в шесть этажей, богатые лавки.* [Н. М. Карамзин. Письма русского путешественника (1793)]

В рассматриваемых сочиненных цепочках наблюдается следующие закономерности:

а) в контекстах до середины 20 в. встречается только кортеж «*здание дом*» [10, 11], в современном языке наоборот — «*дом здание*» [9].

б) в современном языке[6] ясен различительный признак, поскольку он проявляется в половине контекстов, например, в [9]: *дом* — жилой, *здание* — учреждение. До середины 20 в. этой дифференциации, по-видимому, не было, в частности, функция «+учреждение» проявляется только в одном примере и ассоциирована со словом *дом* **[7]**.

О_сочинение_с_обобщением. Контексты находятся для обоих кортежей, но в данной статье рассматривается только <*здание — дом*>. В следующих примерах ракурс — элемент застройки, его оснащение, реконструкция и т. п. Здание является гиперонимом дома:

[12] *Договор строительного подряда заключается на строительство или реконструкцию предприятия, здания (в том числе жилого дома), сооружения или иного объекта, …* [Гражданский кодекс Российской Федерации. Часть вторая (1995)]

[13] *Трудно себе представить современное здание (будь то жилой дом, торговый комплекс или бизнесцентр), которое обходилось бы без лифтов.* [Вверх по рельсам, ведущим вниз (2003) // «Строительство», 2003.05.26]

6.2. Пара слов «дом и строение» (строки 2 и 11)

О_сочинение — 3 примера в ракурсе «ЗВ», где *дом* имеет различительный признак +жилой, который проявляется в виде указания на хозяина дома или проживающих [15, 16]:

[14] *Спаса́ть дома́, строе́ния бесполе́зно — ого́нь распространя́ется с колосса́льной ско́ростью.* [Д. А. Гранин. Месяц вверх ногами (1966)]

[15] *Место сборищ народных было обыкновенно при кирках, возле которых находится всегда несколько строений, дома пастора и кистера, шинок и несколько домов, в которых помещались ремесленники.* [Ф. В. Булгарин. Воспоминания (1846–1849)]

[16] *В следующую ночь дом Белосельских был тоже окружен мушкетерами и пожарными, и в надворных строениях была задержана разбойничья шайка, переселившаяся из дома Гурьева.* [В. А. Гиляровский. Москва и москвичи (1934)]

О_сочинение_с_обобщением. В ракурсе застройки *строение* является гиперонимом *дома*.

[17] *…жили́щный фонд образу́ют находя́щиеся на террито́рии Росси́йской Федера́ции жилы́е дома́, а та́кже жилы́е помеще́ния в други́х строе́ниях (т. е. дома́, строе́ния, строи́тельство кото́рых зако́нчено с осуществле́нием регистра́ции прав на постро́енный объе́кт недви́жи́мости).* [Недостроенное жилье и земельный налог (2004) // «Бухгалтерский учёт», 2004.12.06]

[18] *… возведение каких-либо строений (домов, поселков, предприятий и т. д.* [В. Ларичев и др. Характеристика коррупционных проявлений в сфере природопользования (2 часть) (2004) // «Адвокат», 2004.12.01]

6.3. Пара слов «дом и постройка» (строки 3 и 16)

О_сочинение (8 контекстов). Эта пара встречается исключительно в ракурсах «мини-жил» и «ЗВ», причем *постройки* — во множественном числе. Оба слова часто маркируются по функции, например, *надворные, хозяйственные, совхозные, главный, запасной…* При этом выявляется «прикладной к *дому*» характер *построек* [22, 23], а в [24] отсутствие построек воспринимается как аномалия. Квазисинонимия:

[19] *Не умеют быстро приспосабливать к обороне местные предметы (дома, постройки и т. д.* [Сборник боевых документов Великой Отечественной войны. Выпуск 11 (1941–1945)]

[20] *Все э́ти дни Андрей Сургеев про́жи́л как бы человеком-невиди́мкою, …, а над совхо́зными постро́йками, дома́ми и клу́бом, над маши́нным двором, над совхо́зной землёй, …, над потеря́вшими листву́ дере́вьями — не со́лнце и луна́, не облака́, набу́хшие вла́гой, а ча́вканье и чмо́канье сапожищ Шишлина.* [Анатолий Азольский. Лопушок // «Новый Мир», № 8, 1998]

[21] *Во-вторых, нежелательно, чтобы пруд оказался в тени какой-либо постройки, дома или забора — его обитателям необходим солнечный свет в течение 4–5 часов в день.* [Водный сад (2003) // «Сад своими руками», 2003.03.15]

---

[6] Под «современным языком» иногда понимается русский язык со времени А. С. Пушкина. Для наших слов это — русский язык начиная со второй половины 20 в., когда изменился характер застройки, а с ним и средства языка, ее обслуживающие.

[22] *Когда-то дом был главой целого семейства построек.* [Василий Белов. Плотницкие рассказы (1968)]

[23] *Вместе с другими хозяйственными постройками дом образует маленький готический поселок.* [Николай Гаврюшин. Суханово: соприкосновение двух миров (2002) // «Ландшафтный дизайн», 2002.03.15]

[24] *Седьмой двор выделился издалека какой-то запущенной бедностью — не только потому, что дом был низок и мал и надворные постройки почти отсутствовали, ….* [Сергей Бабаян. Ротмистр Неженцев (1995–1996)]

О_сочинение_с_обобщением. В ракурсах «мини-жил» и «ЗВ» застройка обозначается словом «*постройки*» и ассоциируется со стихийностью (*возникают, разбросанные*) и своеобразием (*стиль*). *Дом* — гипоним *постройки*::

[25] *Одна за другой в «Муромцево» возникают выполненные в едином стиле постройки: дом главный и дом запасной, дом управляющего, летний театр, купальня, охотничий домик.* [Мария Ожерельева. Современники называли ее «царскою» (2002) // «Ландшафтный дизайн», 2002.01.15]

[26] *Все постройки были солидные, кирпичные, дома крыты черепицей, архитектура частью русская.* [В. А. Обручев. От Кяхты до Кульджи. Путешествие в Центральную Азию и Китай (1940)]

6.4. Пара слов «дом и корпус» (строки 4 и 21)

О_. Пара встречается в ракурсах «мини-жил» и «ЗВ». Слова обозначают объекты, которые представлены в дискурсе как разные, причем *корпус* обозначает **часть** *дома*, что определяется выделенными жирным шрифтом определениями к слову *корпус*:

[27] *Все перечисленные системы, агрегаты и механизмы удобно разместить в пристройке или **отдельном от** дома корпусе вместе с гаражами, мастерской и другими необходимыми в усадьбе подсобными помещениями*. [Михаил Харит. Автономный дом (2002) // «Домовой», 2002.05.04]

[28] *Возле дома… горел костер …. Дом был большой, **в несколько** корпусов.* [В. П. Катаев. Отче наш (1946)]

[29] *Основная часть дома была на фундаменте, исключение составлял **боковой** корпус, где помещались кухня, туалет, склад дров и запасная комната, где жил слуга-сарт.* [А. А. Татищев. Земли и люди: В гуще переселенческого движения (1906–1921) (1928)]

6.5. Пара слов «дом и сооружение» (строки 5 и 26)

О_сочинение_с_обобщением. Есть только один пример, диагностирующий родовидовое отношение, которое создается контекстом и «процессной» ассоциацией значения слова *сооружение* (см. р. 2), в следующих примерах слова, поддерживающие эту ассоциацию, выделены жирным шрифтом**:**

[30] *Малой динамической сложностью (неопределенностью) обладают такие **системы**, как **строительные конструкции**, сооружения: дома, мосты, мебель, автомобили, самолеты, …* [И. М. Гуревич. Законы информатики — основа исследований и проектирования сложных систем (2003) // «Информационные технологии», 2003.11.24]

[31] *В ролевой игре по курсу «Основы производственной психологии и педагогики» студенты **возводят** сооружение (дом, школу, гостиницу).* [Вуз будущего // «Студенческий меридиан», 1985]

[32] *Все его сооружения — жилые дома, здания городских учреждений, заводские цеха, проезжие и пешеходные дороги — покошлись на сваях.* [Евгений Рубин. Пан или пропал. Жизнеописание (1999–2000)]

6.6. Пара слов «здание и строение» (строки 7 и 12)

О_сочинение. Всего 2 контекста, не считая штампа «*здания, строения, сооружения*». Следующий пример из текста закона отнесем к терминам (см. р. 5*): «…сооружения связи — объекты инженерной инфраструктуры, в том числе здания, строения, созданные или приспособленные для размещения средств связи и кабелей электросвязи;* [Федеральный закон «О связи» (2003)]», второй пример не принадлежит к современному языку, как и единственный пример, свидетельствующий о синонимии слов здание — строение:

О_ : *До тех пор великие князья московские жили не иначе, как в деревянных домах, да и вообще на всем русском севере каменными зданиями были только церкви: жилые строения были исключительно деревянные.* [Н. И. Костомаров. Русская история в жизнеописаниях ее главнейших деятелей. Выпуск второй: XV-XVI столетия (1862–1875)]. Следователь-

но, в современном языке эти слова вне официально-деловой сферы не ассоциированы друг с другом в контекстах однородности.

6.7. Пара слов «здание и постройка» (строки 8 и 17)

Встречаются в трех видах однородных контекстов, но в иных, чем *дом-постройки* ракурсах. Различительные признаки *построек* — «несамостоятельность», «древность» — поддерживающие контексты выделены жирным шрифтом.

О_сочинение:

[33] *Для разрушения особо прочных зданий, построек и инженерных сооружений создаются корпусные или дивизионные группы разрушения из артиллерии особой мощности и гвардейских миномётных частей (М-31).* [Сборник боевых документов Великой Отечественной войны. Выпуск 1 (1941–1945)]

О_сочинение_с_обобщением:

[34] *Внутри шестиугольной крепости обозначены три постройки: здание с двумя башнями, «кордегардия» («Corp de Guard») и «дом коменданта» («Comend Huus»).* [Т. А. Базарова. План петровского Петербурга. Источниковедческое исследование (2003)].

О_: — квазисинонимия

[35] *При занятии противником **прилегающих к занятому зданию** построек организует их очищение в том же порядке, как указано выше.* [Сборник боевых документов Великой Отечественной войны. Выпуск 1 (1941–1945)]

[36] *Судя по тому, что здания этого времени сохранялись хуже, чем постройки **домонгольской поры**, многое в умениях было утрачено — как правило, уже через 100–150 лет каменные здания приходили в негодность.* [С. А. Еремеева. Лекции по русскому искусству (2000)]

6.8. Пара слов «здание и корпус» (строки 9 и 22)

Встречаются с РКО специального назначения: кадетским корпусом, духовной академией, больницей, комплексом администрации и др. *Корпусом* называется *главный* объект и все другие объекты, связанные со специализированной функцией всего заведения. Здания — тоже функционально отмечены, но не специальными функциями, а обычными (культурными) функциями. Различительный признак «+главн.» реализуется разными прилагательными: *главный корпус*, но *основное здание*.

О_сочинение:

[37] *«Храмы, башни, стены монастыря, здания, корпуса, часовни — все в полуразвалившемся состоянии»* [архимандрит Макарий (Веретенников). Первый наместник возрожденной Лавры // «Альфа и Омега», 2001]

[38] *Пансионат «Крымское приморье» — три **жилых** корпуса; **процедурный** корпус; здание **клуба-столовой** (с колоннами в добротном сталинском стиле),* …[Ольга Ляпунова. Отчет-лоция о путешествии в Крым (2002)].

[39] *Отправленный с рабочим визитом в Грозный новый вице-мэр Москвы Михаил Мень заявил, что параллельно восстановлению **основного** здания будет возводиться и **дополнительный** корпус поврежденной чеченской администрации.* [Нина Зверева. Воспитательная спецоперация Владимира Путина (2003)].

6.9. Пара слов «здание и сооружение» (строки 9 и 22)

О_сочинение_с_обобщением (аналогично р. 6.5):

[40] *…сооружения связи — объекты инженерной инфраструктуры, в том числе здания, строения, созданные или приспособленные для размещения средств связи и кабелей электросвязи*; [Федеральный закон «О связи» (2003)].

6.10. Пара слов «строение и корпус» (строки 14 и 23)

На эту пару слов встретилось всего три примера. Два — на кортеж *корпус — строение* относятся к 19 в. и производят впечатление устаревших: а) *Он заложил еще новый сад и новый корпус, строение для дворовых.* [Л. Н. Толстой. Война и мир. Том третий (1867–1869)]; б) *Та половина была на дворе длинная пристройка к главному корпусу строения.* [Ф. Ф. Вигель. Записки (1850–1860)], так как в слово *строение* без определений в современном языке (см. сноску 7) не ассоциируется с жилым объектом в а) и не может обозначать главный объект минизастройки в б). Пример на кортеж *строение — корпус* **[1]** в р. 1 относится к типу **О_** со сравнительным предикатом. Мог бы свидетельствовать о синони-

мии, но скорее это квазисинонимия с различительным признаком корпуса «+главный», который в тексте и фигурирует.

6.11. Пара слов «постройка и сооружение» (строки 20 и 29)

Только сочинительный контекст и всего два примера. Выделенные жирным шрифтом контексты указывают на маргинальность объектов и привязанность к месту.

О_сочинение:

[41] *возводить на полученных в установленном порядке земельных участках* ***постоянные или временные*** *постройки, сооружения и дороги, необходимые для осуществления хозяйственной деятельности, связанной с пользованием животным миром;* [Федеральный закон «О животном мире» (1995)]

[42] *Здания, сооружения, постройки давно* ***физически устарели****, большинство приемных пунктов рыбы еще колхозной постройки, капитальные и текущие ремонты их не производились десятки лет.* [Строить быстро экономично, качественно (1976) // «Маяк севера», 1976.09.04].

7. Выводы

Предложен Метод корпусного исследования ЛСО между двумя словами и изложены результаты исследования ЛСО внутри пар слов из одной семантической группы: *дом, здание, строение, постройка, корпус, сооружение* этим методом на материале НКРЯ (8).

Метод состоит в а) отборе из корпуса контекстов совместного употребления пары изучаемых слов; б) классификации контекстов на однородные (слова обозначают разные объекты) и тождественные (слова обозначают один и тот же объект); в) классификации синтаксических конструкций рассматриваемого контекста на предмет возможности определенных синтаксических конструкций диагностировать определенные ЛСО. В отличие от психолингвистических диагностических контекстов[7] такие реальные контексты диагностируют ЛСО с учетом комплексных условий конкретного речевого акта, получивших в статье название «ракурсы».

Результаты получены с учетом следующих ограничений: а) метод корпусного исследования ЛСО, примененный к парам слов; б) рассматривались контексты из НКРЯ не ранее второй половины 20 в., поскольку до этого времени застройка и средства языка ее отражающие, были иными; в) за недостатком места представлены результаты исследования только однородных контекстов.

Рассматривались три типа отношений: синонимия, квазисинонимия, родовидовое. Показано, что в основе двух последних лежит понятие различительного признака. Для рассматриваемой группы слов с указанными ограничениями оказалось важным выделить следующие различительные признаки: «+жилой», «+учреждение», «+специальный», «+хозяин», «+функциональн.» (например, *хозяйственный, надворный*,…), «+главн.», а также «+несамостоятельность», «+древность», не исключаются и другие. Физические и конструкционные свойства, например, наличие архитектурного замысла, материал, размер и др. не существенны при установлении ЛСО между словами этой группы.

В ходе исследования понадобились следующие четыре ракурса:

1 «застройка», **2** «мини-жил», **3** «мини-спец», **4** «ЗВ» (см. р. 3), которые оказались важными для данной группы слов. Далее приводится сводная таблица типов ЛСО, выявленных для данной группы слов, с комментариями для каждой пары слов.

Дом — здание. Эти слова являются основным средством современного русского языка для обозначения капитальных объектов в рассматриваемой группе слов. Они вступают в ЛСО друг с другом и со всеми другими словами рассматриваемой группы (исключение составляет пара *здание* и *строение*). Эти слова вступают в ЛСО в ракурсах **1** и **4**, где они являются синонимами, не вступают в мини-ракурсах. Кроме того, в ракурсе **1** *здание* является гиперонимом *дома*, который имеет различительный признак «+жилой». В ракурсе **4** *дом* и *здание* являются квазисинонимами с различительными признаками «+жилой» и «+учреждение», соответственно.

Строение — дом. В ракурсе **1** *строение* — гипероним *дома*, в ракурсе **4** они — квазисинонимы, причем дом имеет различительный признак «+хозяин (проживающий)».

Постройка (мн. ч-ло) — *дом, здание*. В мини-ракурсах *постройки* являются гиперонимом *дома* и *здания*, причем *дом* употребляется в **2**, а *здание* — в **3** ракурсе. *Постройки* также являются квазисинонимом *дома* в двух ракурсах, причем в ракурсе **2** маркированным членом является *дом* с различительным признаком «+главн.», а в ракурсе **3** маркированным членом являются *постройки* с различительным признаком «+функциональн.». Наконец, в ракурсе **1** *постройки* являются маркированным квазисинонимом *здание* с различительным признаком «+несамостоятельноть», «+древность».

[7] Фактически психолингвистические контексты рассматривают ЛСО только в одном из возможных ракурсов, остается вопрос, можно ли рассматривать этот ракурс как обобщение всех других.

Табл. 3. ЛСО для слов *дом, здание, строение, постройка, корпус, сооружение* в контекстах однородности

		1 застройка	**2** мини-жил	**3** мини-спец	**4** ЗВ	
1	дом	↔; ←+жилой|			↔ +жилой|+учрежден.	здание
2	дом	←			+ хозяин|_	строение
3	дом		← +главн|?		← ?|+функциональн.	постройки (мн.)
6	здание	?|+несам., древность		←		постройки (мн.)
4	дом, здание	←|+результат процесса				сооружение
5	здание			0|+спец		корпус
6	—часть→		дом(ед.)	здание	дом (ед.)	корпус
7	строение			0|+глав		корпус
8	постройки				?|?	сооружения

Сооружение — дом, здание. Сооружение является гиперонимом и квазисинонимом слов дом и здание и имеет различительный признак «+результат процесса», обозначая любой результат созидательной («соорудительной») деятельности в соответствии с определением в (9).

Корпус — дом, здание. Здание и *корпус* встречаются только в ракурсе **3**, причем *корпус* является или **частью** здания или его квазисинонимом, маркированным признаком «+спец.». *Корпус* и *дом* встречаются в ракурсах **2** и **4**, причем корпус находится в отношении «**часть**» к дому. *Корпус* не используется в ракурсе **1**.

Строение и *постройка* являются точными синонимами и разделяются по сфере употребления: в официально-деловой сфере используется слово *строение*, *постройка* — крайне редко. Поскольку совместного употребления в одном примере этих слов не встретилось, синонимия этих слов не указана в сводной табл. 3.

Строение — корпус встречаются в одном примере **[1]** в ракурсе **3** как квазисинонимы, причем *корпус* маркирован признаком «+главный».

Постройка — сооружение — квазисинонимы в ракурсах **3–4**. В имеющихся двух примерах речь идет о хозяйственных объектах.

Приведенной таблицей можно пользоваться для генерации синтаксических конструкций, соответствующих указанным отношениям и для номинации объектов в определенных ракурсах. В работах (1) и (2) мы уже использовали понятия, уточняющие особенности дискурного присутствия объекта, называя их «когнитивные перспективы». Это более узкое, чем ракурс понятие, позволяющее как бы «поворачивать» объект разными его ипостасями, например, *озеро* как точка и как зеркало. Когнитивные перспективы влияют на выбор глагола для реализации локативных отношений на изображении. В данной статье мы исследовали понятие ракурса, связанное с номинацией объектов. Не очевидно место ракурса в модели генерации текста, оно может быть связано как с онтологией ситуаций, так и со словарем, обеспечивающим обозначение понятий словами естественного языка.

8. Направления дальнейших исследований

Проведенное исследование является только первым шагом в изучении ЛСО слов с помощью корпусов текстов и требует продолжения для проверки и углубленного исследования сформулированных выше результатов. Требует изучения понятие «ракурс». Необходимо сравнить результаты исследования в контекстах однородности с результатами в контекстах тождества. Требуется проверка по корпусу описаний изображений (см. в (3)) полученных результатов и выработка правил употребления средств языка для описания объектов на изображении. Продолжения и более тщательной разработки требует классификация синтаксических конструкций. Следует принимать во внимание, что в процессе порождения речи участвуют многие факторы, поэтому правила генерации не могут быть простыми. Возможно, из результатов исследований, подобных данному, можно выделить однофакторные правила, но это требует серьезных исследований.

Литература

1. *Болдасов М., Соколова Е.* Онтология и планирование в генерации описаний изображений // Вопросы ИИ, № 2 (в печати)
2. *Boldasov M. V. Sokolova E. G.* Ontology for image annotating // 17th International Conference on Conceptual Structures ICCS'09. Knowledge and ontology "ELSWARE-2009" (Moscow 26–31 July, 2009). Russia, M.: University higher school of economics, 2009. P. 13–21.
3. *Соколова Е. Г.* Об использовании семантических отношений для описания изображений // Вестник РГГУ №8/07 Серия «Языкознание», М.: Издательский Центр РГГУ, 2007. с. 131–144.
4. *Соколова Е. Г., Болдасов М. В.* Формализованное описание содержания изображения как данные для генерации текста // Компьютерная лингвистика и интеллектуальные технологии. Труды Международной конференции Диалог'2007, «Бекасово», 30 мая — 3 июня 2007, М.: Издательский Центр РГГУ, 2007. с. 508–515.
5. *Новый* объяснительный словарь синонимов русского языка (под общ. ред. Ю. Д. Апресяна). Первый выпуск. М: «Школа "Языки русской культуры"», 1999.
6. *WordNet* http://wordnetweb.princeton.edu/perl/webwn
7. *Словарь* синонимов русского языка в двух томах (главн. ред. А. П. Евгеньева). Л.: «Наука», 1970, 1971.
8. *Национальный* корпус русского языка: www.ruscorpora.ru
9. *Малый* академический словарь — Словарь русского языка в 4-х томах // Электронный ресурс ФЭБ. — [М.]: 2005. http://feb-web.ru/feb/mas/mas-abc/0encyc.htm

Речевые нарушения при близкородственном билингвизме: опыт корпусного исследования спонтанной белорусской речи[1]

Speech disfluencies in bilinguals speaking close cognate languages: a corpus study of spontaneous belarusian speech

Сомин А. А. (somin@tut.by)

Российский государственный гуманитарный университет, Москва

На материале корпуса устных текстов исследуются речевые сбои и другие виды нарушений, вызванных интерференцией при русско-белорусском билингвизме. Имеющиеся в литературе данные о стратегиях преодоления речевых сбоев в русском языке сравниваются с полученнымиданными о сбоях при билингвизме.

1. Введение

1.1. Данная работа выполнена в рамках коллективного проекта, направленного на создание экспериментального многоязычного корпуса устных рассказов по картинкам с их последующим пересказом (см. подробнее [Хуршудян 2006] о русской, английской, итальянской и армянской частях этого корпуса). В настоящей работе используется собранный и размеченный автором подкорпус рассказов на белорусском языке.

1.2. Известно, что основным языком, используемым в Беларуси, является русский. Белорусский язык обязательно изучается в школах, и бóльшая часть русскоговорящих является пассивными носителями белорусского языка. Однако имеется определённое число белорусов, которые по тем или иным причинам переходят с русского языка на белорусский в качестве основного. В связи со сменой языка уже во взрослом возрасте речь таких носителей в большей или меньшей степени подвержена интерференции. Рассказы именно таких носителей составили наш корпус.

1.3. Случайный выбор информантов с разным стажем белорусскоязычной жизни, разным уровнем владения языком и разным количеством сфер его использования позволил сформировать корпус, сбалансированный с точки зрения интерферированности речи. Речь людей, более свободно владеющих белорусским, позволила нам получить общую информацию об оформлении речевых сбоев в белорусском дискурсе; речь же людей, хуже овладевших белорусским языком, позволила изучить особый тип нарушений идеального речепорождения (ideal delivery, о понятии см. [Clark & Clark 1977]), связанных с интерференцией: очевидно, что нарушения идеального речепорождения — это наиболее чувствительные к билингвизму точки. В работе рассматриваются как особенности оформления речевых сбоев в белорусском языке в целом, так и отдельно нарушения, связанные с интерференцией.

1.4. Следует заметить, что до последнего времени основные работы, посвящённые билингвизму в Беларуси, описывали два основных вида белорусско-русской интерференции: проявления белорусских элементов в русской речи и так называемую классическую трасянку (смешанный белорусско-русский код, вызванный родным белорусским языком) (ср., например, [Михневич ред. 1985], [Бірыла, Супрун ред. 1982] и др.). Интерес к русско-белорусской интерференции возник сравнительно недавно, и количество работ, посвящённых ей, невелико. Из последних работ следует выделить диссертацию [Малько 2004], описывающую лексическую и грамматическую интереференцию в белорусских текстах печатных СМИ. Очевидно, что единственно возможным методом для аналогичного исследования устной белорусской речи является корпусный.

1.5. Наш корпус представлен двумя основными частями — аудиозаписями и их транскриптами. Он состоит из 40 текстов общим объёмом примерно

[1] Исследование выполнено при поддержке гранта РФФИ 10-06-00338.

9000 словоупотреблений (включая обрывы слов) или 2150 элементарных дискурсивных единиц (далее ЭДЕ; ЭДЕ — это минимальный квант устного дискурса, приблизительно соответствующий интонационно цельнооформленной клаузе; о понятии см. [Кибрик, Подлесская ред. 2009]). Транскрибирование речи осуществлялось по методике, предложенной в вышеупомянутой работе, с учётом специальных решений, принятых для белорусского языка. Общая продолжительность звучащей речи составляет около 75 минут.

2. Нарушения идеального речепорождения

В нашей работе мы опишем речевые сбои и затруднения (как связанные с билингвизмом, так и свойственные Языку вообще, но имеющие в белорусском какое-то особое проявление), а также прочие нарушения идеального речепорождения, связанные с актуализированным двуязычием, которые не были замечены говорящими и потому не повлекшие за собой речевые сбои.

2.1. Речевые сбои и затруднения

2.1.1. Речевые сбои свойственны устной речи на любом языке мира, они неизбежны. Способы оформления речевых сбоев и стратегии их исправления варьируются от языка к языку. В начале исследования мы выдвинули гипотезу, согласно которой в устной белорусской речи билингвов будут использованы сегментные маркеры хезитации, свойственные русской речи. Впоследствии эта гипотеза подтвердилась практически полностью, однако мы не можем с уверенностью сказать, были ли эти маркеры заимствованы из русского языка или же развились в белорусской речи самостоятельно. Заметим, что простые короткие маркеры полностью идентичны русским, а более сложные и длинные совпадают с ними с точностью до естественных различий в фонетике, грамматике и лексике русского и белорусского языков (вось ‘вот’, значыцца ‘значится’, уласна кажучы ‘собственно говоря’, як бы ‘как бы’ и др.):

(43) Даш_R2[2]

66.8	42	...(0.8) \Вось.	вот
68.0	43	..(0.1) /І значыцца /–пакатаўся,	и значится покатался
69.3	44	....(1.1) {ЦОКАНЬЕ 0.1} вельмі /–задаволены застаўсяw,	очень доволен остался

(44) Бар_R1

39.1	24	..(0.4) /\Пайшоў-ў ён ..(0.2) у-у аўтамабільны /–салон,	пошёл он в автомобильный салон
41.9	25	....(1.2) мм(0.3) /–ну і ўласна кажучы прыйш= || /–сказаў:	ну и собственно говоря прид= || сказал

(45) Бар_T2

11.5	7	..(0.4) ммм(0.9) ..(0.4) /\ну \перш за ўсёw ...(0.5) як бы вырашыў /падсілкавацца трохі,	ну прежде всего как бы решил подкрепиться немного

Несмотря на практически полную идентичность маркеров в двух языках, употребление некоторых маркеров в белорусском может отличаться от аналогичного в русском. Например, маркер ну может находиться в начале высказывания, начинающегося с союза але ‘но’:

(46) Кац_T1

106.3	47	..(0.3) ''(0.2) /–дзеці ...(0.7) н-не /зразумелі прыкола_	дети не поняли прикола
109.4	48	...(0.9) ''(0.4) ..(0.3) /–ну але мабыць у /\наступны раз ён ..(0.2) падорыць ёй –сапраўдную {CMEX 0.6} машыну {CMEX 1.2}.	ну но может в следующий раз он подарит ей настоящую машину

Заметим также, что после але может находиться и частица ж ‘же’, что тоже невозможно в русском языке (таким образом, союз але по сочетаемости похож на русский союз однако):

[2] Здесь и далее в заголовке примера указывается код того рассказа в нашем корпусе, из которого взят данный пример. В начале каждой строки примера указывается время ее начала в секундах (отсчитывая от абсолютного начала рассказа), а затем номер данной ЭДЕ в рамках рассказа. Числа в скобках обозначают длительность пауз в секундах. Строка сопровождается дословным переводом на русский язык. При необходимости строка может сопровождаться неформальным комментарием транскрайбера, для этого используется пятая, дополнительная колонка транскрипта. О других конвенциях дискурсивной транскрипции см. [Кибрик и Подлесская (ред.) 2009].

(47) Алй_R1

92.9	53	...(0.9) I-i || ээ(0.4) ''(0.3) ...(0.9) {ЦОКАНЬЕ 0.1} ..(0.4) але ж-ж ..(0.1) трэба было неяк ..(0.1) *выкручвацца* з /\сітуацыі тя= || ''(0.3) *складанай,*	и но же нужно было как-то выкручиваться из ситуации тя= || тяжёлой	тя= — интерферема

Ещё одним отличием в употреблении маркеров являются особенности маркера неяк ‘как-то’: в тех контекстах, где неяк употребляется в качестве аппроксиматора («сигнала о том, что говорящий не несёт полной ответственности за ту номинацию, которой предшествует маркер, так как она в некотором роде неполна или не совсем подходящим образом описывает ситуацию» [Podlesskaya (in print); *перевод мой — А. С.*]), в русском языке употребление маркера как-то маловероятно. Скорее, маркер неяк соответствует русскому как-то вот или там:

(48) Алй_T2

29.6	13	'''(0.8) ...(0.5) Ён-н \д-доўга ..(0.4) \н-неяк маа= || меў /намер,	он долго как-то ма= || имел намерение	возможно, ма= употреблено как инициаль основы презенса глагола мець ‘иметь’
34.3	14	..(0.3) гэта –зрабіць,	это сделать	

(49) Зьм_R2

124.6	69	*Потым яны дасталі нейкага* смажанага /–кабанчыка яшчэ_	потом они достали какого-то жареного кабанчика ещё
127.0	70	...(0.5) {CMEX 1.2} ..(0.4) каб неяк –пад’есці троху,	чтобы как-то перекусить немного

(50) Маш_T2

5.5	6	..(0.3) ''(0.3) /Ён неяк з= || /–звычайна з’ядаў свой ..(0.2) /–сняданак_	он как-то об=|| обычно съедал свой завтрак

Ещё одной особенностью выражения хезитации в белорусском языке является отсутствие при хезитации сандхи на границе слов, в противоположность ожидаемому при непаузированной речи. По правилам орфоэпии звуки [u] и [i], находящиеся в начале слова, в контексте предыдущей гласной переходят в [w] и [j] соответственно, напр. ён увайшоў ‘он вошёл’ ([u]), но яна ўвайшла ‘она вошла’ ([w]). Однако если говорящий удлиняет соответствующую фонему, выражая свою хезитацию, данное чередование не происходит:

(51) Кац_T1

21.2	9	....(1.2) /пайшоў ён спачатку-у у-у /жаночую ..(0.3) к-краму,	пошёл он сначала в женский магазин

(вместо ожидаемого спачатку ў жаночую);

(52) Ул_R1

54.3	20	не вельмі с= || не /спадабаўся,	не очень п= || не понравился
55.1	21	/\і-і –Змітру \перапала з гэтага.	и Змитру попало за это

(вместо ожидаемого спадабаўся й Змітру).

Сандхи также не происходит при выражении хезитации в виде незаполненной паузы:

(53) Акс_T1

1.3	2	які ..(0.3) у адну цудоўную /раніцу ўсп= || ..(0.3) /прыгадаў,	который в одно чудесное утро всп= || вспомнил

При выражении хезитации в виде заполненной паузы, наоборот, происходит не ожидаемое сандхи:

(54) Янк_T1

39.3	22	/–але ж-ж ..(0.1) ээ(0.2) ў /\выніку яна йдзе ў /аўтасалон_	но же в результате она идёт в автосалон

2.1.2. Перейдём к **речевым сбоям, связанным с билингвизмом и интерференцией**. В нашем случае такие нарушения похожи на сбои, возникающие при овладении иностранным языком.

2.1.2.1. Хезитации

а) Часто хезитационные паузы могут возникать при поиске говорящим забытого или неизвестного ему слова, при том, что он помнит это слово на русском языке:

(55) Ста_Т2

99.0	43	і-іі (0.9) ..{ЦОКАНЬЕ}..(1.7) ’’’(0.4) ...(0.5) калі-іw ..(0.1) пачуў ..(0.3) кошт ....(1.0) самай таннай /машыны,	и когда услышал стоимость самой дешёвой машины	
106.0	44	ён быў вельмі ==	он был очень ==	
107.0	45	....(1.2) ммм(0.5) ....(2.3) ’’(0.2) ....(1.7) ён быў ==	он был ==	
113.3	46	..(0.2) ён ..(0.1) вельмі ....(1.0) так –скажам ==	он очень так скажем	ён со смехом так скажам очень тихо
115.8	48	..(0.3) {СМЕХ 0.4} ....(2.0) /\А!	а	
118.7	49	..(0.4) ээ(0.3) ..(0.3) *Калі ён пачуў* ..(0.2) –кошт ..(0.2) ээ(0.3) ’’(0.2) /машыны,	когда он услышал стоимость машины	
122.1	50	..(0.3) самай \таннай,	самой дешёвой	
123.1	51	..(0.4) ’’’(0.7) ..(0.3) ён –вырашыў ...(0.5) сы́сці з / кірмаша,	он решил уйти с рынка	неверное ударение в сысці́

В приведённом выше примере говорящий (как он впоследствии сообщил) пытался вспомнить белорусский аналог слова поник. Хезитация при поиске забытого слова продолжалась на протяжении 3 ЭДЕ (12,5 секунд), при этом говорящий прибегал к заполненным и незаполненным паузам, частичным повторам клаузы и маркеру поиска так скажам, необходимом для «оттягивания времени» после затянувшейся паузы и сохранения статуса «я ещё говорю». После этого говорящий сообразил, как выйти из сложившейся ситуации (о чём свидетельствует ЭДЕ 48 и повтор ЭДЕ 43, хоть и не точный, в строчках 49–50).

Приведём другой пример. Говорящая калькирует русское слово времяпрепровождение (белорусский аналог которого переводные словари дают описательными сочетаниями баўленне часу или правядзенне часу) и, переведя первую его часть, задумывается над переводом второй:

(56) Але_R2

17.9	9	..(0.1) пасля \такога вот ээ(0.1) часу= ..(0.3) =\право↑джання,	после такого вот время= =препровождения

Это характерный пример интерференции: лексема часаправоджанне (в написании через а; хотя вариант, произнесённый информанткой, является более верным) в интернете встречается всего лишь три раза.

В целом следует отметить, что многие говорящие в процессе речи изменяют уже начатую конструкцию, чтобы избежать употребления незнакомого или забытого слова или же слова, в принадлежности которого к белорусскому (а не русскому) языку они не уверены.

б) Перешедшим с русского языка на белорусский свойственен языковой пуризм, который в частности проявляется в желании употреблять слова, отличающиеся от их русских переводов. При наличии двух белорусских синонимов один из них может оцениваться как русизм, и тогда говорящий может колебаться при выборе желаемого:

(57) Акс_Т1

1.3	2	які ..(0.3) у адну цудоўную /раніцу усп= || ..(0.3) /прыгадаў,	который в одно чудесное утро всп= || вспомнил

(58) Алй_R1

17.0	10	што б ..(0.3) такое ім \падарыць.	что бы такое им подарить
18.7	11	...(0.7) Падарава= || ...(0.5) па= ==	подари= || по= ==
20.4	12	\так.	так.

В первом случае говорящая из двух синонимов успомніць и прыгадаць выбирает второй, как не совпадающий с русским словом. Во втором случае информант сомневается в выборе из двух белорусских аналогов русского подарить — падарыць и падараваць, но, так и не выбрав, решает, что адресату для понимания будет достаточно уже произнесённого им, и возвращается к повествованию. Причину его

сомнения можно объяснить тем, что падарыць некоторыми информантами осознаётся как русизм, из-за чего они стараются избегать этого глагола в речи.

в) Обращает на себя внимание тот факт, что говорящие, совершив ошибку, вызванную интерференцией, стремятся скорее её исправить, не заостряя на ней внимание, и как бы «зачёркивают» предыдущее сказанное слово, не употребляя при этом никаких речевых маркеров. В частности, маркеры ой и ой-ой в нашем корпусе встретились только после смысловых ошибок (т.е. связанных с планом содержания), и ни разу после собственно языковых ошибок (т.е. связанных с планом выражения).

2.1.2.2. Исправления

Вслед за [Кибрик, Подлесская (ред.) 2009], мы выделяем два типа исправлений: (он-лайн) коррекции (или собственно **коррекцию**) и рестроспективную корреккцию (или **редактирование**, см. пункт *г)*).

а) «В режиме **коррекции** говорящий реагирует на обнаруженную проблему немедленно» [Кибрик, Подлесская ред. 2009].

Одним из типов коррекции является **модификация.** «Говорящий использует модификацию в тех случаях, когда ревизии подлежит не исходное смысловое задание, а его конкретное воплощение. В этих случаях забракованный фрагмент корректируется с полным или частичным сохранением значения» [Кибрик, Подлесская ред. 2009]. Особым типом модификаций следует считать **модификации**, **вызванные употреблением русизма**.

Влияние интерференции «выливается» в модификацию в том случае, когда говорящий начинает произносить русское слово, замечает это, обрывает слово и замещает его белорусским аналогом, спасая всю ЭДЕ «малой кровью»:

(59) Бар_R1

0.0	1	ээ(0.3) Прыбл= || ээ(0.4) /\набліжаўся Новы /–год,	прибл= || приближался Новый год

(60) Алй_R1

92.9	55	…(0.9) І-і || ээ(0.4) ’’(0.3) …(0.9) {ЦОКАНЬЕ 0.1} ..(0.4) але ж-ж ..(0.1) трэба было неяк ..(0.1) *выкручвацца* з /\сітуацыі тя= || ’’(0.3) *складанай,*	и но же нужно было как-то выкручиваться из ситуации тя= || тяжёлой

б) Аналогично, особым видом модификации нужно считать и самоисправления в случае употребления двух синонимов, один из которых оценивается говорящим как русизм, см. выше (15)–(16), а также следующий пример:

(61) Ул_R1

15.0	7	..(0.4) ’’(0.3) і раш= || ээ(0.2) /–вырашыў звярнуцца-а да сваіх \дзяцей.	и реш= || решил обратиться к своим детям

Заметим, что глагол рашыць часть словарей считает абсолютным аналогом русского решить и синонимом глагола вырашыць, а другая часть ограничивает сферу его употребления только значением ‘решить (задачу)’, оставляя все остальные значения глаголу вырашыць.

в) Весьма характерен следующий пример коррекции, связанной со сложностью образования формы косвенного падежа при чередовании конечного согласного основы:

(62) Надз_R1

154.2	97	…(0.8) /\набыў ім па /–цац= || …(0.5) па /–цаццы,	купил им по игруш= || по игрушке

Чередование к/ц в дательном падеже существительного женского рода в сочетании с исходом корня на -ц (цацка ‘игрушка’) приводит к затруднениям у говорящей: она произносит корень и вынуждена оборвать себя, чтобы сообразить, как поступить с суффиксом (очевидно, что чередование в цацка/цаццы гораздо более необычно и непривычно, чем, например, в ночка/ночцы).

К этому случаю примыкают и другие сбои, связанные с отличиями белорусской морфологии от русской. Ср. следующий пример, где сбой также происходит в связи с чередованием заднеязычных при склонении:

(63) Даш_R2

107.9	62	..(0.3) а я проста магу т= || ..(0.1) і на \адной наг= =зе,	а я просто могу т= || и на одной ног= =ге

Конечно, подобного рода чередования являются потенциальным «ошибкоопасным» местом, и, теоретически, причиной ошибок такого рода может быть вовсе и не интерференция. Однако наличие сравнительно большого числа случаев, где палатализация и вовсе не была произведена, на наш взгляд, является косвенным доказательством ведущей роли именно билингвизма в появлении в речи таких ошибок.

г) Боязнь ошибиться в процессе белорусской речи может привести к употреблению гиперкорректных форм. В случае, если говорящий замечает за собой такую ошибку, осмыслив её, он может попробовать ретроспективно её исправить. Такие случаи ретроспективной коррекции мы называем **редактированием** (подробнее см. [Кибрик, Подлесская ред. 2009]): говорящий уже постфактум информирует слушающего о том, что сказанный отрезок подлежит исправлению или уточнению.

Ср. следующий пример:

(64) Ул_T2

55.8	25	...(0.5) /\Ну aw пасля /–гэтага яму-у нека= эээ(0.5) || ..(0.2) больш за /паў-ўгоду прыйшлося хадзіць у \гіпсу.	ну а после этого ему неко= || больше полугода пришлось ходить в гипсу.
60.6	26	....(1.1) \Гі↑псе.	гипсе

Здесь говорящий употребил гиперкорректную форму предложного падежа гіпсу (распределение окончаний предложного падежа -е/-у в русском и белорусском языках не совпадает, что служит частым источником ошибок в белорусской речи). Больше секунды ему потребовалось на осознание ошибки, после чего он выполнил редактирование (характерно интонационное выделение исправленного окончания — подъём тона на безударном окончании).

д) Конечно, как уже упоминалось, не во всех случаях речевых сбоев можно с уверенностью говорить об интерференциальных причинах их появления. Иногда приходится ставить интерференцию в один ряд и с другими «кандидатами» на причину самоисправления. Ср., например, следующий случай:

(65) Алс_R1

77.9	37	....(1.4) ээ(0.2) не –аднойчы ...(0.7) яму гэты “–Ягуар” ==	не раз ему этот “Ягуар” ==
82.0	38	...(0.8) не аднойчы ён /сніў гэты “Ягуар” у сваіх ...(0.7) \найлепшых /снах,	не раз он видел (*досл.* снил) этот “Ягуар” в своих лучших снах

В белорусском языке допустимы оба варианта глагола, имеющего значение ‘видеть во сне’ — сніць и безличное сніцца. Но не исключено, что говорящий решил заменить предполагаемый сніцца на сніць, потому что фраза с невозвратным глаголом звучит «более по-белорусски» (то есть «менее по-русски», ср. замечание о языковом пуризме). Хотя, разумеется, этот речевой сбой мог быть вызван и другими причинами, к примеру, невозможностью употребления запланированного фрагмента у сваіх найлепшых снах после безличного глагола сніцца.

2.2. Не замеченные говорящими нарушения, связанные с билингвизмом.

2.2.1. Во многих случаях говорящие не замечают сделанных в речи ошибок. Таким образом, эти ошибки не приводят к речевым сбоям, хотя, разумеется, являются нарушениями идеального речепорождения. Подобные нарушения могут происходить на всех уровнях языка. Приведём примеры лексической, морфологической и синтаксической интерференции:

(66) Алс_T2 (слова от разных корней)

183.0	66	Замест ...(0.8) /лыжаў яму прыйшлося хадзіць з \кастылямі.	вместо лыж ему пришлось ходить с костылями	Интерферема. Надо мыліцамі.

(67) Ста_T1 (паралексы — межъязыковые паронимы)

132.5	54	...(0.5) /Ён набыў ..(0.2) проста дзіцячую /–машыну,	он купил просто детскую машину	
134.8	55	...(0.5) ’’(0.3) –калекцыённую,	колекционную	Интерферема. Надо калекцыйную.

(68) Бар_R1 (ударение и паралексы)

78.6	38	..(0.3) /\ён ...(0.9) падараваў сваей-й /–жонцы ..(0.1) /–ма́ленькую /–мадэльку ..(0.3) \ **а́ўтамабіля**.	он подарил своей жене маленькую модельку автомобиля	Интерферема. Надо сваёй. Интерферема. Надо мале́нькую.

(69) Зьм_R2 (вопросительные/относительные местоимения)

14.8	8	што як раз /сёння /ён /збіраўся /наведаць ...(0.6) сваіх старых /\сяброў,	что как раз сегодня он собирался навестить своих старых друзей	
18.1	9	каторых ужо *даўным-даўно* не /бачыў,	которых уже давным-давно не видел	Интерферема. Надо якіх (т. к. придаточное)

(70) Ста_T1 (морфология: палатализация заднеязычных)

10.4	6	...(0.5) ''(0.2) што падараваць сваёй –жон↑ке ... (0.7) на-а дзень \народзінаў.	что подарить своей жене на день рождения	Интерферема. Надо жонцы

(71) Ула_T2 (синтаксис)

3.4	3	ў \семь гадзін /раніцы,	в семь часов утра	Интерферема. Надо а сёмай гадзіне.

Впрочем, вариант обозначения времени с предлогом у — не вполне очевидный русизм: это встречается и в литературном белорусском языке, так что, возможно, его не следует считать нарушением нормы.

Приведём также примеры на такое нарушение идеального речепорождения, как смешение паронимов:

(72) Алй_T2

121.8	54	..(0.2) і /яго нецвярозы /–стан ...(0.5) ''(0.3) не стаў /дапаможнікам,	и его нетрезвое состояние не стало пособием	Смешение паронимов

(73) Але_R2

12.6	7	..(0.3) з /якімі /–ё-ён ...(0.5) /\праводзіў ..(0.4) / вільготны \час.	с которыми он проводил влажное время	Смешение паронимов

В первом случае происходит смешение паронимов дапаможнік ‘пособие’ и памочнік ‘помощник’, причиной чему служит то, что гораздо более частым аналогом русского помочь является белорусское дапамагчы с приставкой. Во втором случае говорящая путает слова вільготны ‘влажный’ и вальготны, которое в некоторых контекстах может значить ‘свободный’. Впрочем, даже верное употребление не помогло бы говорящей избежать ошибки, так как белорусским аналогом выражения свободное время является вольны час.

2.2.2. Впрочем, не все русизмы стоит считать нарушением идеального речепорождения. В некоторых случаях говорящие специально включают в свою речь русизмы для создания опеределённого эффекта, чаще всего комического или иронического.

(74) Кас_T2

28.9	27	\/–выпілі /\вінца-а_	выпили винца	
30.2	28	..(0.1) /яшчо раз вінца,	ещё раз винца	Интерферема. Надо яшчэ
31.4	29	і /–зноў вінца выпілі_	и снова винца выпили	

Интересно, что аналогичное явление, только зеркально отражённое, встречается в русском языке русскоязычного населения Беларуси (см. примеры в работе [Лисковец 2006].

Иногда употребление русизма можно объяснить наличием у русской лексемы определённых коннотаций, отсутствующих у её белорусского аналога:

(75) Кас_R1

50.5	39	...(0.8) /А \дочанька кажа	а доченька говорит	
52.0	40	«/Не папаша@	«нет папаша	интерферема

2.2.1. Фонетика

Если в работе [Михневич ред. 1985] речь идёт о сильном влиянии белорусской фонетики на русскую речь белорусов, то в речи носителей, участвовавших в записи нашего корпуса, наблюдается сильное влияние русского языка. Речь идёт о мягких [rj] и [tɕ] (в нормативном белорусском языке [r] и [tʂ] только твёрдые), «иканьи» вместо «яканья», а также

сильной редукции гласных, несвойственной белорусскому языку. Фонетическое несоответствие нормам также является нарушением идеального речепорождения.

3. Перспективы дальнейшего исследования

3.1. При создании корпуса от каждого носителя записывалось два рассказа на белорусском языке и два рассказа на русском. Представляет интерес сравнение в будущем, с одной стороны, русского языка белорусскоязычных билингвов и русских монолингвов (мы предполагаем использовать русский подкорпус многоязычного корпуса рассказов по картинкам собранный и описанный в работе [Хуршудян 2006]), а с другой стороны — русского и белорусского языка билингвов.

Нами выдвинуто две гипотезы:

а) в белорусской речи билингвов количество сбоев будет бóльшим, нежели в их русской речи (в связи с худшей усвоенностью белорусского языка);

б) речевые сбои в русской речи билингвов и монолингвов не будут существенно различаться.

Проверка этих гипотез — дело будущих исследований.

3.2. Представляет интерес изучение устной диалектальной белорусской речи, особенно в Полесском регионе и на белорусско-польском пограничье (здесь могут наблюдаться стратегии, характерные соответственно украинской и польской речи), а также изучение особенностей устной дискурса на трасянке.

3.3. Просодически размеченный корпус со столь подробной разметкой может быть полезен при изучении многих разнообразных аспектов белорусской речи. Также рассматривается возможность включения данного корпуса в состав создаваемого в данный момент Национального корпуса белорусского языка.

Литература

1. *Бірыла М. В., Супрун А. Я.* (рэд.) Пытанні білінгвізму і ўзаемадзеяння моў: Зб. // Мінск: Навука і тэхніка, 1982.
2. *Кибрик А. А., Подлесская В. И.* (ред.) Рассказы о сновидениях: Корпусное исследование устного русского дискурса. // М.: ЯСК, 2009.
3. *Лисковец И. В.* Русский и белорусский языки в Минске: проблемы билингвизма и отношения к языку. Дис. ... канд. филол. наук. // СПб: ЕУСПб, 2006.
4. *Малько Г. І.* Руска-беларуская інтэрферэнцыя ў перыядычным друку Рэспублікі Беларусь (лексічны і граматычны ўзроўні). Дыс. ... канд. філал. навук. Мінск: БДУ, 2004.
5. *Михневич А. Е.* (ред.) Русский язык в Белоруссии. // Минск: Наука и техника, 1985.
6. *Хуршудян В. Г.* Средства выражения хезитации в устном армянском дискурсе в типологической перспективе. Дис. ... канд. филол. наук. // М.: РГГУ, 2006.
7. *Clark H. H., Clark E. V.* Psychology and Language // New York: Harcourt Brace Jovanovich, 1977.
8. *Podlesskaya V. I.* (in print) Parameters for Typological Variation of Placeholders // [Typological Studies in Languages Series]. Fillers, pauses, and placeholders. Ed. by N. Amiridze, B. Davis, M. Maclagan. J. Benjamins.

Синтаксический анализатор «Treevial». Принцип динамического ранжирования гипотез[1]

«Treevial» syntax parser. Paradigm of the dynamic hypothesis ranking

Старостин А. С. (starost@rinet.ru),
Арефьев Н. В. (nick.arefyev@gmail.com),
Мальковский М. Г. (malk@cs.msu.su)

Московский государственный университет им. М. В. Ломоносова

В статье подробно обсуждается главный принцип, в соответствии с которым работает синтаксический анализатор Treevial, — *принцип динамического ранжирования гипотез*. Описывается формальный аппарат, используемый в Treevial для описания грамматики. Отдельный раздел посвящен механизму штрафов, функционирующему вместе с базовым формализмом. Излагается схема работы анализатора. Обсуждаются достоинства и недостатки предлагаемого подхода. В конце статьи приводится описание инструментов штрафования, заложенных в анализатор.

Введение

Работа посвящена описанию синтаксического анализатора Treevial. Этот анализатор является базовой составляющей системы морфо-синтаксического анализа Treeton (см. [Мальковский, Старостин, 2006]), разрабатываемой на факультете ВМиК МГУ с 2005 г. Система была задумана как исследовательская платформа, базируясь на которой студенты и аспиранты факультета могли бы решать различные задачи компьютерной лингвистики. В данный момент система довольно динамично развивается. В рамках Treeton ведутся исследования в области автоматического морфологического и синтаксического анализа, а также в области автоматизации базовых процедур лингвостиховедческого анализа. Система является некоммерческой и открытой для всех, кто желает принять участие в ее развитии.

Главной темой данной статьи является принцип динамического ранжирования гипотез, заложенный в основу анализатора Treevial. Поясним, в чем он заключается. Пусть имеет место процесс восходящего синтаксического анализа, в рамках которого возникают различные конфигурации, связывающие группы слов друг с другом. Кроме того, допустим, что существует способ некоторым образом оценивать «качество» (степень правдоподобности) каждой возникающей конфигурации. Тогда принцип динамического ранжирования может быть сформулирован следующим образом: **предпочтение более правдоподобных конфигураций менее правдоподобным должно отдаваться не после, а в процессе синтаксического анализа**.

Авторы продемонстрируют, какие достоинства имеет синтаксический анализатор, построенный в соответствии с принципом динамического ранжирования, и какие ограничения приходится накладывать на компоненты анализатора. Статья состоит из четырех разделов. В первом разделе описывается базовый формализм, используемый в Treevial для описания грамматики (способ формулировки синтаксических правил). Во втором разделе обсуждается концепция штрафов. В третьем разделе приводится схема работы анализатора и обсуждается принцип динамического ранжирования. В четвертом разделе описываются инструменты штрафования, которые в данный момент поддерживает анализатор.

1. Формализм для описания грамматики

Язык Treevial позволяет задавать грамматику естественного языка в виде системы декларативных правил, описывающих возможные пути сборки синтаксических структур из исходных словоформ (вариантов морфологического анализа).

[1] Работа выполнена при поддержке РФФИ (проект 08-06-00192)

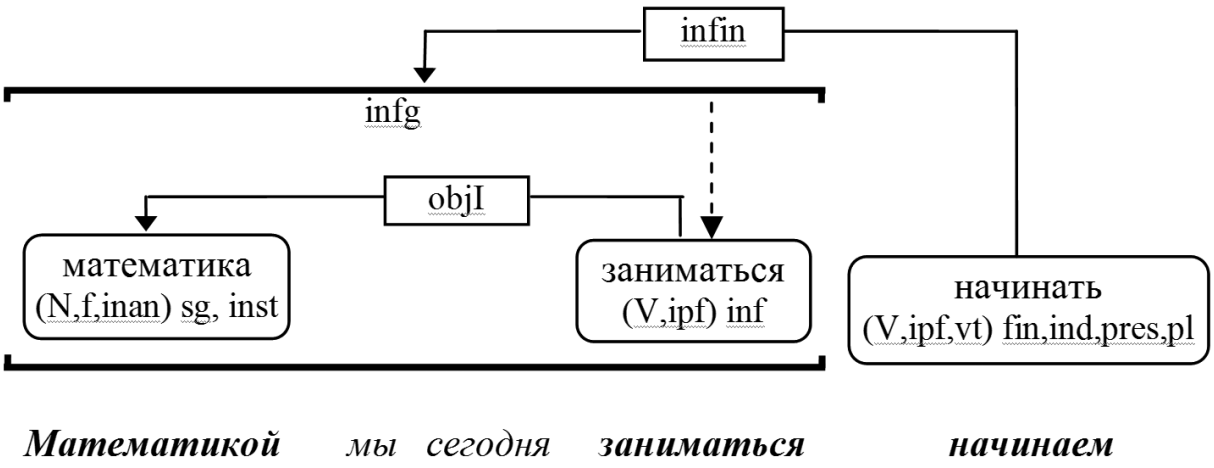


Рис. 1. Пример синтаксической интерпретации

Идеология системы подобна той, которая принимается при работе с категориальными грамматиками — перед анализатором ставится задача для заданной цепочки атрибутированных словоформ (категорий) найти возможность связать все словоформы в одну древовидную структуру (сократить все категории), т. е. доказать принадлежность данной цепочки описываемому языку. Следует отметить, что возможностей организовать одну и ту же цепочку форм в структуру обычно находится больше, чем одна[2]. Наиболее близким к описываемому формальным аппаратом из области категориальных грамматик является [Dekhtyar, Dikovsky, 2008] (в нем допускаются дистантные зависимости между элементами).

Базовым для языка Treevial является понятие синтаксической структуры или синтаксической интерпретации фрагмента предложения. Последний термин предпочтительнее, т. к. он подчеркивает «гипотетичность» всего, что делает анализатор, и имплицирует наличие альтернативных интерпретаций для одного и того же фрагмента. Синтаксическая интерпретация — это направленное дерево с размеченными дугами, узлами которого являются либо морфологические интерпретации слов предложения, либо виртуальные узлы, созданные в процессе анализа. Все элементы синтаксической структуры имеют линейные координаты, т. е. интерпретация всегда спроецирована на исходное предложение. Следует отметить, что такая проекция не обязательно представляет собой сплошной отрезок — допускаются разрывные интерпретации.

На рисунке 1 приводится пример синтаксической интерпретации фрагмента предложения «математикой мы сегодня заниматься начинаем», иллюстрирующий все описательные возможности этого формального объекта: фиксация связей между элементами, объединение элементов в группы (создание виртуальных узлов), потенциальная разрывность. Жирным выделены слова, образующие проекцию синтаксической интерпретации.

Из описанных в литературе способов описания синтаксических структур наиболее близким аналогом синтаксической интерпретации является размеченная система синтаксических групп (РССГ, [Гладкий, 1985]). Возможность создания виртуальных узлов позволяет описывать синтаксические явления в духе грамматик составляющих (ср. [Хомский, 1962]). Возможность использования связей, напротив, допускает взгляд на синтаксические структуры, близкий к грамматикам зависимостей (ср. [Tesnière, 1959; Mel'cuk, 1988]). Существенно, что имеется возможность одновременного использования преимуществ и той, и другой парадигмы.

Язык Treevial задает правила преобразования синтаксических интерпретаций. Каждое правило описывает ситуацию, в которой две интерпретации, обладающие определенными свойствами, и находящиеся в определенных отношениях (например, контактность или согласованность вершин по какой-то категории), могут образовать новую интерпретацию. Например, может быть проведена связь от некоторого элемента одной интерпретации к корню другой или обе интерпретации могут быть объединены в группу, которая получит некоторые новые свойства (и в том, и в другом случае будет образована новая интерпретация). Помимо использования бинарных правил (работающих с двумя интерпретациями), в языке Treevial есть возможность

[2] Заметим, что сходство с категориальными грамматиками на уровне идеологии и заканчивается. Дело в том, что, в категориальных грамматиках центральной идеей является сокращение категорий, т. е. превращение двух элементов (обычно соседних) в один. После того как сокращение было произведено, сокращенный элемент перестает принимать участие в деривации. В языке Treevial подход другой — если два элемента объединяются (образуют группу или связываются связью), то образуется новый составной объект, который в дальнейшем может участвовать в деривации, присоединяя что-либо к любой из своих частей.

задавать унарные правила. Эти правила применяются только к одной интерпретации, обладающей определенными свойствами. Результатом применения унарного правила всегда является создание группы (виртуального узла), обладающей определенными свойствами и включающей исходную интерпретацию.

Схематично синтаксическое правило может быть записано следующим образом:

```
rule MyRule() {
  спецификация аргументов
  ::
  ограничения
  →
  действия
  ::
  условия взимания штрафов
}
```

Действия, описанные в блоке действий, выполняются над теми интерпретациями, узлы которых удовлетворили спецификации аргументов и ограничениям, описанным в блоке ограничений. В результате применения правила создается новая синтаксическая интерпретация. Рассмотрим составные части правила по отдельности.

1.1. Спецификация аргументов

Спецификация аргументов состоит из одного или двух шаблонов (логических выражений, составленных из элементарных ограничений на значение атрибутов, скобок, знаков дизъюнкции и конъюнкции). С шаблонами сопоставляются узлы синтаксических интерпретаций (как корневые, так и промежуточные). Если шаблонов два, то между ними может стоять знак $\sim$. Это означает, что элементы, сопоставляемые с шаблонами могут быть расположены дистантно. Если между двумя шаблонами стоит знак +, это означает, что сопоставляемые с шаблонами элементы должны примыкать друг к другу. При этом каждый из шаблонов может быть окружен квадратными скобками, что означает, что примыкание требуется не от самого элемента, сопоставляемого с шаблоном, а от проекции всего поддерева данного элемента. По умолчанию порядок, в котором должны стоять сопоставляемые с шаблонами элементы, может быть произвольным. Однако, его можно зафиксировать, поставив после спецификации аргументов знак ^.

Следует отметить, что в блоках, описываемых ниже (ограничений, действий, условий взимания штрафов), узел, сопоставленный с левым шаблоном спецификации аргументов, всегда обозначается как *A*, а узел, сопоставленный с правым шаблоном, как *B*.

1.2. Ограничения

Ограничения задаются с помощью логического выражения, в котором могут участвовать атрибуты двух элементов, за счет чего можно учитывать такие явления как, например, согласование. После сопоставления узлов синтаксических интерпретаций и шаблонов из спецификации аргументов переменные *A* и *B* получают конкретные значения и становится возможным проверить, обращается ли логическое выражение в истину или нет.

Пример ограничения (согласование по падежу и числу):

$$A.CAS == B.CAS \;\&\&\; A.NMB == B.NMB$$

1.3. Действия

Блок действий представляет собой список элементарных действий, состоящий минимум из одного элемента. Действия выполняются в случае, если узлы синтаксических структур, сопоставленные с шаблонами, удовлетворяют ограничениям. Вторым необходимым условием для выполнения действия является его корректность по отношению к принципу древесности. Это означает, что ни одно правило не может создать интерпретацию, в которой будут циклы, или у какого-нибудь узла будет более одного хозяина.

Действия бывают четырех типов:

- Связывание

(**имя_шаблона_1,имя_шаблона_2**){**тип_связи**}

Проводится связь определенного типа от узла, сопоставленного с первым шаблоном, к узлу, сопоставленному со вторым шаблоном. Например, *(A,B){myRelType}*.

- Унарная агрегация

C[**имя_шаблона**]{**последовательность присваиваний атрибутов**}

Узел, сопоставленный с шаблоном, помещается внутрь нового элемента, атрибуты которого задаются в результате выполнения последовательности присваиваний (см. ниже). Например, *C[A]{NMB=A.NMB;GTYPE="Clause";CAS=null;}*.

- Бинарная агрегация

C[**имя_шаблона_1,имя_шаблона_2**]
{**последовательность присваиваний атрибутов**}

Узлы, сопоставленные с шаблонами, помещаются внутрь нового элемента «на равных правах», не подчиняясь друг другу. Атрибуты нового узла задаются в результате выполнения последовательности присваиваний. Например, *C[A,B]{CAS=A.CAS;NMB=B.NMB;GTYPE="conj";}*.

- Включение

*имя_шаблона_1**[**имя_шаблона_2**]***

Узел, сопоставленный с шаблоном 2, помещается внутрь элемента, сопоставленного с шаблоном 1. Например, *B[A]*.

При описании бинарной и унарной агрегации упоминалась последовательность присваиваний атрибутов. Это список элементов вида: *название_атрибута = значение.*

В момент выполнения агрегации список проходится слева направо, вычисляются соответствующие значения и выполняются присваивания. Значения могут задаваться как явно (конкретными литералами), так и ссылкой на значение некоторого атрибута одного из узлов, сопоставленных с шаблонами из спецификации аргументов. В качестве значения может также выступать служебное слово *null*. В этом случае при выполнении присваивания соответствующий атрибут стирается.

1.4. Условия взимания штрафов

Условия взимания штрафов очень похожи на ограничения. Однако если при невыполнении ограничения применение правила запрещается, то при невыполнении условия взимания штрафов запрета не происходит — просто штраф получившейся интерпретации увеличивается на некоторую векторную величину (штрафы в системе представляются целочисленными векторами). Подробно штрафы обсуждаются в следующем разделе. Приведем пример условия взимания штрафов:

A.@start < B.@start : (0,0,100)

Это условие означает, что если A начинается левее, чем B, следует увеличить штраф на *(0,0,100)*.

До сих пор не было упомянуто еще одно важное ограничение на процесс применения правил: проекции интерпретаций, участвующих в правилах, не могут пересекаться. Напомним, что проекции синтаксических интерпретаций в общем случае могут быть разрывными. Сформулированное условие гарантирует отсутствие циклов после применения правил и не позволяет использовать в рамках одной структуры различные морфологические трактовки одного и того же слова.

Таким образом, можно резюмировать, что каждое синтаксическое правило применяется к одной или к двум синтаксическим интерпретациям. В каждой интерпретации выбираются узлы, с опорой на которые сначала проверяются ограничения, потом выполняются действия и, наконец, проверяются условия взимания штрафов. Заметим, что благодаря принципу древесности и тому факту, что правило всегда содержит хотя бы одно действие, в качестве опорного узла одной из интерпретаций всегда выбирается ее корень. Другими словами, как минимум одна синтаксическая интерпретация всегда оказывается подчиняемой. А вот для интерпретации, которая играет роль хозяина, опорный узел может выбираться различными способами. Это означает, что могут существовать различные способы применения одного и того же правила к одной и той же комбинации синтаксических интерпретаций. Например, одна и та же предложная группа может быть присоединена к любому из узлов какого-либо «большого» синтаксического дерева, если только этот узел имеет характеристики глагола.

Дополнительно следует сказать, что могут существовать варианты получения одной и той же синтаксической интерпретации различными путями (рис. 2). Это не приводит к размножению гипотез и не влияет существенным образом на эффективность анализатора — система умеет находить дубликаты. Может показаться, что описанный принцип «сборки» синтаксических структур явля-

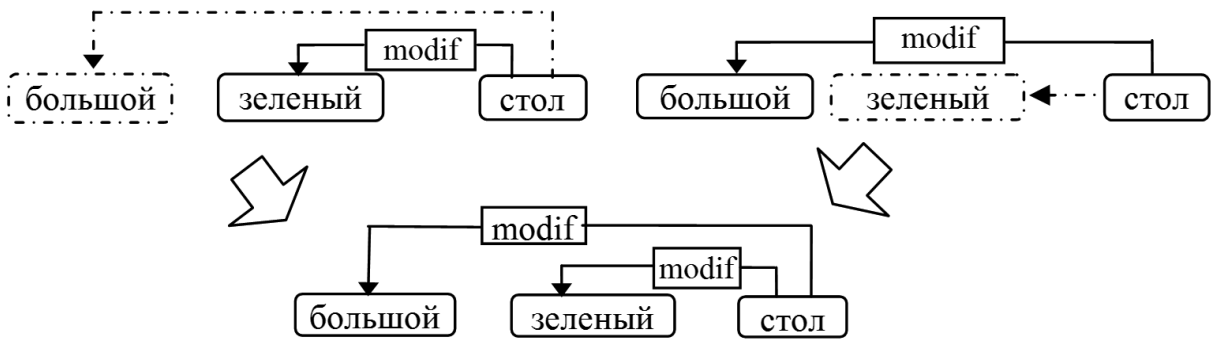


Рис. 2. Пример альтернативных вариантов получения одной структуры

w_1=в; w_2=стиле; w_3=отца;

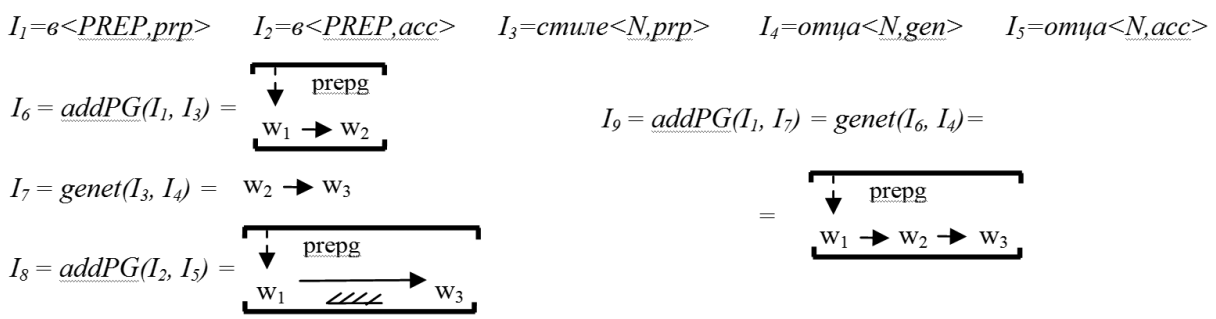


Рис. 3. Пример действия синтаксических правил для словосочетания «В стиле отца»

ется слишком свободным. Однако, любые попытки авторов ограничить деривацию с помощью дополнительного правила (аксиомы) на уровне формального аппарата приводили либо к невозможности описывать сложные языковые явления, либо к проблемам с вычислительной эффективностью[3]. Кроме того, оказалось, что такая «плазмоподобная» модель хорошо согласуется с идеологией штрафов, о которой пойдет речь в следующем разделе.

Еще один аспект предлагаемого аппарата, который следует прокомментировать, это локальность опорных шаблонов. Дело в том, что в рамках синтаксического правила есть возможность обращаться только к одному узлу каждой из интерпретаций (тому, который сопоставился с шаблоном). Это значит, что мы неявно исходим из предпосылки, что все свойства необходимые для формулировки правила будут сконцентрированы в одном узле. Однако, практика показывает, что в естественных языках встречаются так называемые дальние согласования (ср. [Тестелец, 2001]). В этих случаях для формулировки корректного ограничения нужно обращаться не к самому узлу, а к одному из его потомков (*видишь **таблицу**, один из столбцов **которой** заполнен нулями?,* жирным выделены согласующиеся слова). В языке Treevial существует специальная конструкция, позволяющая учитывать дальние согласования. Ее описание выходит за рамки данной статьи.

На рисунке 3 демонстрируется работа анализатора с системой из двух синтаксических правил. Первое правило создает предложную группу, второе устанавливает посессивное (или генитивное) отношение:

```
rule addPG {
  {POS == "PREP"}+[{POS == "N"}] ^
  ::
  A.GCAS == B.CAS
  →
  (A,B){preposit}
  C[A]{PHRASE="prepg";}
}

rule genet {
  {POS == "N"}~{POS == "N" && CAS ==
  "gen"}
  →
  (A,B){genet}
}
```

2. Аппарат штрафов

Прежде чем приступить к описанию аппарата штрафов, остановимся на нескольких моментах, обуславливающих необходимость его использования.

Во-первых, легко видеть, что несмотря на то, что в синтаксических правилах разрешено учитывать геометрически дистантные зависимости, правила, по сути своей, являются контекстно-свободными, т. к. отсутствует возможность учитывать какие-то факты, касающиеся частей уже собранного синтаксического дерева, не имеющих отношения к опорным узлам. Например, нельзя написать правило, выражающее следующий смысл: «такое-то слово можно связать с таким-то, если между ними (не) стоит некоторое слово». По нашему глубокому убеждению, описание синтаксиса естественного языка может быть построено без использования контекстно-зависимых правил. Более того, предлагаемый в этом разделе аппарат штрафов в определенном смысле исключает потребность в использовании таких правил.

Второе, о чем следует сказать, это изначально принятая авторами установка на малый объем лингвистических средств. Вслед за [Перцов, Старостин, 1999] мы придерживаемся идеи о необходимости создания синтаксического анализатора, способного опираться в основном на морфологические характеристики слов без использования богатой словарной информации о сочетаемости. Такая установка про-

[3] Проблемы возникали в первую очередь при анализе непроективных предложений.

диктована в первую очередь тем, что в открытом доступе не существует хорошо формализованных словарей сочетаемости единиц достаточно большого объема, в то время как грамматический словарь А. А. Зализняка доступен всем.

Оба упомянутых факта приводят к тому, что системы правил, которые можно написать на языке Treevial, помимо правильного разбора предложения всегда допускают большое количество альтернативных неправильных вариантов анализа. Многие из них даже могут быть осознаны носителем языка при некотором умственном усилии (ему приходится либо распознать какой-то сложный интонационный контур, допускающий странный порядок слов, либо как-то исказить смысл предложения, либо допустить нехарактерное управление и т. п.). Помимо очевидной проблемы выбора правильного варианта анализа, порождение большого количества результатов имеет еще один существенный недостаток — оно происходит медленно. Полный анализ предложения из 15 слов может занимать порядка 10 минут, из которых 99% времени будет потрачено на очень далекие от действительности варианты анализа.

Путь дальнейшего усложнения языка синтаксических правил казался тупиковым. Было принято решение «разделить переменные» и попытаться разработать обособленный механизм, позволяющий ограничивать функционирование базового механизма (аппарата синтаксических правил).

Таким образом возникла идея ранжирования спектра гипотез, порождаемых анализатором, в соответствии с некоторой эвристической функцией, называемой *штрафной функцией*. Штрафная функция соотносит с любой синтаксической интерпретацией неотрицательное действительное число. При вычислении этой функции могут использоваться различные характеристики структуры: уровень проективности, количество повторений неповторимых связей, наличие пропусков слов в структуре, корректность расстановки запятых, условия взимания штрафов в конкретных правилах, согласованность с моделями управления (при наличии последних) и др. В принципе, при ее вычислении могла бы учитываться и семантика (при наличии соответствующих формальных средств). Различные явления, влияющие на значения штрафной функции, описываются набором штрафных конструкций языка Treevial. Примером такой конструкции является упоминавшееся в предыдущем разделе условие взимания штрафов, формулируемое в теле синтаксического правила. В четвертом разделе этой статьи коротко описываются другие возможности языка Treevial, позволяющие управлять вычислением штрафной функции.

Идея сортировки всех вариантов анализа относительно некоторой шкалы качества хороша с логической точки зрения, но, к сожалению, сама по себе она никак не решает проблему производительности. Очевидно, что принцип «сначала получи все, потом вычисли значения штрафной функции и отсортируй» оказывается нежизнеспособен. Встал вопрос о том, можно ли построить алгоритм синтаксического анализа таким образом, чтобы менее штрафованные варианты обнаруживались раньше, чем более штрафованные. Понятно, что выполнение такого требования имеет высокую практическую ценность. Во-первых, результаты анализа можно использовать, не дожидаясь завершения работы алгоритма. Кроме того, появляется возможность отсекать гипотезы в соответствии с некоторым пороговым значением штрафа. Оказалось, что выполнение такого требования становится возможным, если наложить некоторые ограничения на штрафную функцию. Будем считать, что заданы система синтаксических правил R и штрафная функция $P(I)$. Тогда ограничение формулируется следующим образом:

Для любого правила $r \in R$ должно быть верно, что для любых I_i из того, что $I_r = r(I_0, I_1)$, следует, что $P(I_r) \geq P(I_i)$ $(i=0,1)$

Другими словами, штрафная функция должна быть неубывающей (если рассматривать множество интерпретаций как частично упорядоченное[4]). На бытовом языке это можно сформулировать следующим образом: «качество целого никогда не должно превосходить качество составляющих». Ниже будет показано, каким образом приведенное ограничение позволяет построить анализатор, удовлетворяющий требованию динамической сортировки результатов. Но сначала необходимо обсудить, как строится штрафная функция, которая, с одной стороны, удовлетворяет этому ограничению и, с другой стороны, имеет под собой лингвистические основания.

В ранних версиях анализатора Treevial авторы пытались строить требуемую функцию, исходя из свойств всего дерева синтаксической интерпретации. Подсчитывались различные суммы различных характеристик (количество пересечений связей, количество повторяемых связей, количество пропусков слов и т. п.). Существенной проблемой оказалась такая характеристика интерпретации, как количество пропущенных слов. Дело в том, что, с одной стороны, взимание штрафов за пропуски слов кажется вещью совершенно естественной и очень существенно ограничивает перебор, но, с другой стороны, очевидно, что если в двух интерпретациях есть пропуски слов, то в их объединении пропусков может уже не быть (они могут друг друга

[4] На интерпретациях можно ввести следующее отношение частичного порядка: $I > I_0$ тогда и только тогда, когда I может быть получено из I_0 в результате применения последовательности синтаксических правил. Рекурсивно это выражается так: **1)** если $I = r(I_0,I_1)$, то $I > I_0$ и $I > I_1$; **2)** если $I = r(I_m,I_n)$ и либо $I_m > I_0$, либо $I_n > I_0$, то $I > I_0$.

дополнить). Это означает, что с учетом этой характеристики свойство неубывания штрафа не выполняется. Не ограничивать возможности анализатора пропускать слова авторы не могли, т. к. это было критично с точки зрения производительности. Полный запрет на пропуск слов оказался слишком жестким ограничением (этот вопрос дополнительно обсуждается в четвертом разделе). Кроме того, существовала некоторая логическая нестыковка в самой постановке задачи. Требовалось построить функцию, свойства которой формулируются, исходя из предпосылки, что аргумент функции имеет рекурсивную природу, а при этом при построении функции рекурсивность объекта никак не учитывается.

Все это навело авторов на мысль о внесении в систему дополнительного условия (или закона), опирающегося на рекурсивную природу объекта, за счет которого свойство неубывания штрафа может быть строго доказано. Это условие носит название *условия инкрементальности*. Идея очень проста: по определению считается, что штраф результата применения некоторого правила к нескольким интерпретациям равен сумме штрафов этих интерпретаций и некоторой неотрицательной составляющей, величина которой зависит от соединяемых интерпретаций и от действий, выполняемых при применении правила. Штраф конкретной интерпретации при этом вычисляется как наименьший из штрафов, которые могут быть вычислены в соответствии с условием инкрементальности при различных способах получения данной интерпретации. Математически это выражается следующим образом:

$$P(I) = \min_{i,j,r} P(I_i) + P(I_j) + f(I_r),$$
$$\text{где } I = I_r = r(I_i, I_j),\ f \geq 0$$

Таким образом, мы приходим к модели, при которой каждая синтаксическая интерпретация ассоциируется с оптимальным способом ее сборки. Это довольно существенное идеологическое изменение. При таком подходе в ряде случаев становится важным не только то, как устроена синтаксическая интерпретация (какие в ней связи и группы), но и то, как именно она была создана, т. к. в ее штрафную оценку включаются штрафы ее предков. Практика показывает, что такое усложнение не вызывает существенных проблем при работе с анализатором. Оптимальный способ сборки интерпретации сохраняется в памяти анализатора. Поэтому всегда есть возможность его проанализировать и понять, почему величина штрафа интерпретации именно такая. То, как именно описанная идеология влияет на проблему пропуска слов, обсуждается в четвертом разделе данной статьи. Здесь скажем лишь, что при таком подходе предложения, которые не могут быть собраны без пропусков слов, попадают в особую штрафную зону.

Дополнительно следует отметить, что с точки зрения эффективности алгоритма анализа важным является требование того, чтобы и функция f была построена рекурсивно, т. е. для ее вычисления использовались бы какие-то интегральные характеристики соединяемых интерпретаций и то, какие именно действия к ним применяются. Обращение ко всему дереву интерпретации при вычислении этой функции нежелательно. В текущей реализации анализатора Treevial процедура вычисления штрафа для интерпретации построена именно таким образом. В частности, пример штрафов за непроективность, описанных в четвертом разделе, хорошо иллюстрирует сказанное.

До сих пор о штрафах говорилось как о действительных числах. Это позволило доступно изложить концепцию штрафования интерпретаций. Однако, в реальности анализатор устроен немного сложнее. На практике работать с одним действительным числом оказывается неудобно, т. к. возникает естественное желание не смешивать показатели, характеризующие различные свойства структуры. Например, уровень непроективности структур и количество повторений одинаковых связей, исходящих из одного узла, разумно наблюдать как независимые показатели. В связи с этим в Treevial все штрафы представляются векторами из действительных чисел. В управляющем файле анализатора есть возможность назначить каждому из учитываемых свойств свой разряд штрафного вектора. Все изложенное выше, с небольшими изменениями остается верным и для штрафных векторов, т. к. во всех случаях, когда требуется одно действительное число (для сравнения величин штрафов), для штрафного вектора вычисляется норма (равная сумме всех элементов).

Использование штрафных векторов позволяет не только «навести порядок» среди штрафных показателей, но и гибко настраивать анализатор при анализе конкретных текстов. Появляется возможность отсекать структуры, в которых наблюдаются определенные явления. Пусть, например, заранее известно, что в анализируемых текстах не встречаются атрибутивные инверсии (ср. *человек умный*). Тогда, если в системе правил заложено, что в определенном разряде подсчитываются штрафы на инверсии, то при работе анализатора можно отсечь все интерпретации, штрафные вектора которых имеют в этом разряде ненулевые значения.

3. Схема работы анализатора

Как уже говорилось, цель анализатора — найти все возможности «уложить» слова входного предложения в одну синтаксическую структуру или, другими словами, найти все синтаксические интерпретации входного предложения. Каждый вариант

морфологического анализа каждого слова во входном предложении трактуется системой как элементарная синтаксическая интерпретация (состоящая из одного узла). Начинается процесс применения правил. Возникают новые интерпретации, в свою очередь участвующие в правилах. Процесс замыкается. Заканчивается он в тот момент, когда не остается ни одного варианта применения ни одного правила ни к одной комбинации интерпретаций. То, какая именно комбинация интерпретаций должна быть обработана в конкретный момент, определяется по значениям штрафных векторов синтаксических интерпретаций, образующих комбинацию. Система всегда предпочитает комбинации интерпретаций, суммарный штраф которых на текущий момент минимален. В случае наличия нескольких комбинаций с одинаковым суммарным штрафом, предпочтение отдается комбинации, интерпретации которой покрывают наибольшее число слов входного предложения. Интерпретации, проекция которых покрывает все предложение, помещаются в очередь результатов[5]. Одинаковые интерпретации, возникающие в процессе анализа, «схлопываются» — из всех одинаковых интерпретаций остается одна, штраф которой минимален.

В случае, если выполнено свойство неубывания штрафной функции, построенный таким образом процесс гарантирует упорядоченность результатов относительно нее[6].

Для того, чтобы эффективно реализовать описанную переборную схему, потребовалось запрограммировать специальные объекты, называемые *комбинаторами*. Комбинатор представляет собой объект с n входами и одним выходом (в текущей реализации n никогда не превышает 2). Входы комбинатора являются динамически сортируемыми относительно штрафной функции множествами синтаксических интерпретаций. На выходе комбинатор генерирует различные комбинации из n синтаксических интерпретаций. При этом выполняются следующие условия:

1. Элемент комбинации с номером i принадлежит входному потоку с номером i.
2. Комбинации никогда не повторяются.
3. В каждый момент времени на выходе комбинатора находится комбинация с наименьшей из возможных на данный момент суммой штрафов.

Важным свойством комбинатора является то, что он способен динамически реагировать на изменения во входных множествах. Несколько упрощая, можно сказать, что если в каком-то входном множестве появляется «хороший» элемент, то комбинатор автоматически переключается и начинает генерировать комбинации с этим элементом.

Перед началом анализа система синтаксических правил преобразуется в систему комбинаторов. Это преобразование может осуществляться различными способами. Самый простой принцип следующий: с каждым правилом в соответствии с количеством аргументов соотносится набор множеств интерпретаций и один комбинатор[7].

Для того, чтобы на каждом шаге выбиралась оптимальная комбинация, требуется из набора текущих выходов комбинаторов выбирать лучший. Для этого уже сами комбинаторы организуются в список, который динамически поддерживается в упорядоченном состоянии.

Для того, чтобы возникающие синтаксические структуры эффективно распределялись по входным потокам комбинаторов, используется специальный модуль, динамически «индексирующий» синтаксические интерпретации.

---

[5] В действительности анализатор проверяет не только полноту покрытия, но и соответствие корня интерпретации некоторому шаблону, задаваемому в управляющем файле. Этот механизм описывается в четвертом разделе статьи.

[6] На самом деле мы сознательно упрощаем описание. Для того, чтобы упорядоченность результатов была действительно гарантирована, необходимо наличие промежуточного буфера, в который попадают результаты. На рисунке 4 этот буфер обозначен.

[7] В Treevial на сегодняшний день используется более сложная схема, связанная с использованием знаков препинания. Алгоритм анализа построен так, что предложение перед началом обработки делится на минимальные фрагменты, ограничиваемые знаками препинания. Анализ сначала происходит внутри этих фрагментов. Если интерпретация достигает границы фрагмента (соприкасается с делимитатором, ограничивающим фрагмент), она получает возможность комбинироваться с интерпретациями из соседнего фрагмента. В четвертом разделе описывается механизм штрафования, контролирующий «соприкосновения» интерпретаций с делимитаторами. Таким образом, при анализе предложения из n фрагментов в памяти анализатора возникает n элементарных систем комбинаторов (построенных по принципу «одно правило — один комбинатор»).

На рисунке 4 приводится схема, иллюстрирующая процесс анализа. Эта схема реализует принцип динамического ранжирования гипотез. Остановимся на основных достоинствах и недостатках предлагаемого подхода.

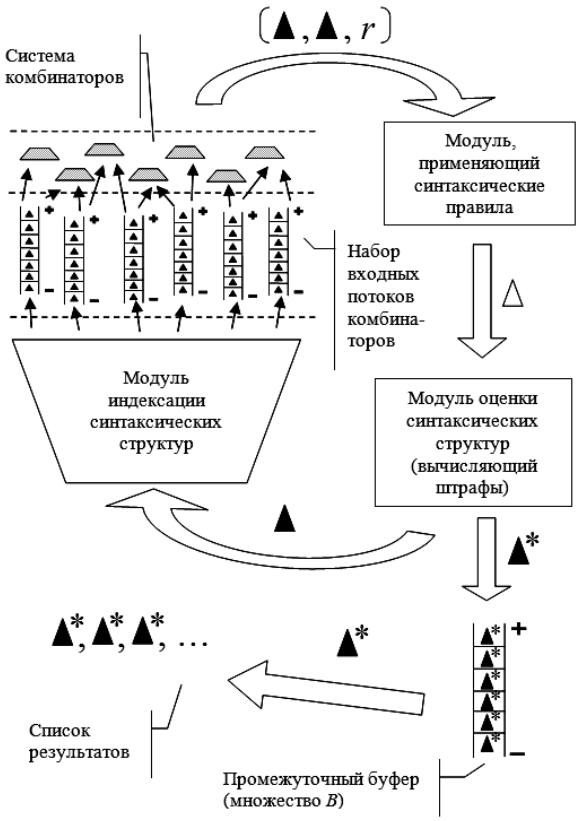


Рис. 4. Схема работы синтаксического анализатора. Треугольники обозначают синтаксические интерпретации. Прозрачные треугольники обозначают интерпретации, для которых еще не посчитан штраф. Треугольники со звездочкой обозначают интерпретации, покрывающие все входное предложение.

Основным достоинством описанной схемы является то, что она позволяет использовать для записи синтаксических правил формальный аппарат, предоставляющий высокую степень свободы. Например, в нем есть возможность не требовать контактности от соединяемых элементов. С помощью этого формализма удается лаконично описывать синтаксис на некотором базовом уровне, но без штрафов его использование невозможно, т. к. количество порождаемых вариантов для любого предложения средней длины оказывается огромным. С помощью аппарата штрафов, в свою очередь, удается моделировать ограничения, накладываемые на базовый механизм. Схема динамического ранжирования позволяет совместить две модели так, чтобы они работали параллельно. За счет того, что оценка интерпретаций происходит в процессе, а не после анализа, наименее штрафованные варианты обнаруживаются очень быстро, т. к. благодаря неубыванию штрафов «хорошие» интерпретации по определению образуются от «хороших». В нормальной ситуации, если штрафной механизм адекватно настроен, то те варианты, которые анализатор порождает первыми, оказываются ближе всего к действительности. Поэтому, получив, к примеру, один или пять первых результатов (в зависимости от решаемой задачи), процесс можно останавливать. С помощью такого приема вычислительные ресурсы экономятся очень существенно.

Предлагаемая парадигма оказывается очень удобной при управлении процессом синтаксического анализа. За счет того, что штрафы за каждое из учитываемых явлений накапливаются в определенных разрядах, появляется возможность влиять на процесс анализа, различными способами задавая допустимые значения для каждого из разрядов. Например, становится возможным осуществлять анализ в несколько этапов. При первом проходе анализатор можно запускать с большим количеством ограничений (запрещая разрывы, непроективность, повторимость и т. п.). Если при этом удается получить хороший результат, то анализ можно не продолжать. Если же результатов нет или они слишком штрафованы (по тем параметрам, которые не были ограничены), можно продолжить анализ, прямо «на лету» расширив зону допустимых штрафов. Такой подход дает существенный выигрыш в производительности, т. к. первый запуск оказывается молниеносным и в большом количестве случаев позволяет получить результат.

С точки зрения программной архитектуры, схема динамического ранжирования также имеет некоторые достоинства. Тот факт, что модуль оценки синтаксических структур отделен от остальных частей системы, позволяет легко добавлять в систему новые механизмы штрафования, в рамках которых могут использоваться произвольные алгоритмы. К примеру, все механизмы, описывающиеся в четвертом разделе, добавлялись в систему постепенно и каждое добавление не производило «революции» в программном коде. Хорошим примером, иллюстрирующим то, как отсутствие инкапсулированности оценочного механизма может вызывать проблемы, являются вероятностные контекстно-свободные грамматики (PCFG). Из [Johnson, 1998] хорошо видно, что жесткая привязка вероятностных оценок к продукциям контекстно-свободной грамматики (их аналогом являются правила Treevial) не позволяет в полной мере учитывать совместную встречаемость единиц. Следует отметить, что в Treevial на данный момент существует лишь эскизный вариант механизма штрафов на лексическую сочетаемость, который не может конкурировать с PCFG по количеству сил и времени, потраченных исследователями. Однако установка на учет контекстных зависимостей заложена в этот

механизм изначально. Во многом за счет того, что правила и модуль оценки логически разделены, это оказалось легко сделать.

Отдельно следует сказать, что в рамках описанной выше концепции достаточно четкие очертания обретает схема взаимодействия со всей системой абстрактного модуля, моделирующего семантику. Причем, не имеет значения, как именно этот модуль будет устроен внутри. Важно, что он должен функционировать параллельно с описанной переборной схемой. В динамике для каждой синтаксической интерпретации могли бы строится семантические представления. Если при этом они оценивались бы аналогично синтаксическим, то семантические штрафы влияли бы на ход анализа наряду с синтаксическими. Существенным кажется то, что при таком подходе «обретать смысл» будут не только конечные структуры , но и все фрагментарные. Это в чем-то согласуется с человеческим механизмом восприятия — мы начинаем осмыслять то, что слышим, сразу, не дожидаясь конца сообщения.

Авторы столкнулись с двумя недостатками схемы динамического ранжирования. Первый из них связан с требованием неубывания штрафной функции и с условием инкрементальности. Одним из существенных следствий этих постулатов является невозможность взимания штрафов за отсутствие каких либо элементов структуры. Так, например, если у слова есть какая-то обязательная (или очень желательная) валентность, невозможно взять штраф за то, что она не выражена в анализируемом предложении. Точнее сказать, такие штрафы невозможно взимать в динамике (после завершения анализа целую структуру всегда можно проанализировать и оценить каким угодно образом). Попытка взимать такие штрафы при анализе фрагментарных интерпретаций в сочетании с принципом инкрементальности неизбежно приведет к противоречивой ситуации. Так, например, если в предложении присутствует слово, заполняющее обязательную валентность, то интерпретации, в которых это слово еще не присоединено к хозяину, будут оштрафованы (за незаполненную валентность), а интерпретации, в которых валентность, наконец, заполнится, окажутся оштрафованными «незаслуженно» (т. к. вторые породятся из первых, а штрафы накапливаются). Существуют технические приемы, позволяющие обойти это ограничение, но все они имеют недостатки. Следует отметить, что на практике без использования штрафов за отсутствие удается обходиться. Для авторов вопрос о необходимости использования штрафов этого типа до сих пор остается открытым.

Вторым недостатком описанной схемы является необходимость соотносить друг с другом абсолютные величины штрафов, отражающих различные явления. На сегодняшний день это делается вручную при настройке анализатора на корпус текстов и представляет собой достаточно трудоемкий процесс. В системе Treeton существует графическая среда, позволяющая производить синтаксический анализ коллекции предложений, при работе с которой процесс балансировки штрафов несколько упрощается: система автоматически отслеживает динамику изменений штрафных векторов конкретных предложений при изменении системы штрафов и правил.

В ближайшее время авторы не планируют предпринимать шаги в направлении автоматизации процесса балансировки штрафов, т. к. первоочередной задачей на данный момент является расширение системы правил и штрафов для русского языка и тестирование анализатора на корпусе. Только после того, как это будет сделано, станет ясно, какое количество штрафных разрядов требуется для качественного анализа. Необходимость использования методов автоматической настройки существенным образом зависит от порядка этой величины.

4. Инструменты штрафования

В этом разделе описываются средства языка Treevial, позволяющие влиять на значения штрафной функции в процессе анализа. Мы сознательно приводим лишь концептуальное описание этих инструментов, т. к. полноценное описание, снабженное примерами и фрагментами формального синтаксиса, заняло бы место, сопоставимое по размеру со всей статьей.

4.1. Штрафы за непроективность

В анализаторе Treevial предусмотрены возможности для оценки уровня проективности структур. В момент применения каждого правила системой вычисляются проекции соединяемых узлов. Если между проекциями есть зазор (одно или более слов), взимается штраф за непроективность (его величина настраивается в управляющем файле). Помимо базового механизма в Treevial поддерживается возможность тонкой настройки политики взимания штрафов за непроективность, позволяющая учитывать то, какие синтаксические фрагменты являются статистически более плотными (неразрывными), а какие допускают разрывы. Примером первых могут служить отдельные конъюнкты, входящие в сочинительную группу, вынос слов из которых едва ли возможен (ср. **умный пришли Вася и глупая Даша*). Примером фрагмента, допускающего разрывы, может служить инфинитивный оборот (ср. *кожаную ты хочешь купить куртку?*). Тонкий контроль проективности в языке Treevial обеспечивается с помощью специальной таблицы из шаблонов и штрафных векторов.

4.2. Штрафы за повторение связей

Штрафы за повторение используются в тех случаях, когда требуется ограничить количество связей одного типа, выходящих из одного узла синтаксической структуры. В качестве примера можно привести связь между глаголом и существительным в именительном падеже. В русском языке эта связь больше одного раза не повторяется (ср. **Вася Петя играл в футбол*). В момент применения каждого правила проверяется, не была ли добавлена связь, тип которой объявлен неповторимым, к узлу, из которого связь такого типа уже исходит. Если была, то взимается штраф за повторение связей. Список неповторимых связей и величина штрафа задаются в управляющем файле.

4.3. Механизм работы со знаками пунктуации

В анализатор Treevial заложено понятие знака пунктуации (или делимитатора). В управляющем файле имеется конструкция, позволяющая задать шаблон, используя который система может отличить знаки пунктуации от остальных символов. В предыдущем разделе упоминалось, что знаки пунктуации имеют значение для алгоритма анализа, т. к. по ним производится первоначальное разбиение предложения на фрагменты. Для того, чтобы в процессе синтаксического анализа появилась возможность комбинировать некоторую синтаксическую интерпретацию с интерпретацией из соседнего фрагмента, необходимо, чтобы знак пунктуации, стоящий на границе, оказался «синтаксически оправдан» в контексте одной из интерпретаций. Это вычисляется, исходя из специальной таблицы, задаваемой в управляющем файле. Если оправдать знак не удается, доступ все равно открывается, но при этом взимается особый штраф на «неоправданный знак препинания». Такой подход позволяет достаточно гибко работать с пунктуацией. В частности, система оказывается устойчива к текстам, в которых есть пунктуационные ошибки, и к текстам, с нестандартной — несовременной или индивидуально-авторской — пунктуацией.

4.4. Штрафы при применении правил

Штрафы при применении правил уже упоминались в первом разделе. Они используются в тех случаях, когда требуется наложить штраф на то или иное явление, связанное со спецификой конкретного синтаксического правила. Например, штраф на обратный порядок существительного и прилагательного при образовании атрибутивной связи между ними («*человек умный*» вместо «*умный человек*»).

4.5. Штрафы за пропуски

Как уже говорилось, формальный аппарат синтаксических правил допускает создание интерпретаций, проекции которых не являются непрерывными. Однако при анализе реальных предложений необходимость использования разрывных интерпретаций возникает совсем не часто. Заметим, что необходимость использования разрывных интерпретаций в процессе создания и непроективность являются связанными, но отнюдь не эквивалентными свойствами синтаксических интерпретаций. Так, например, предложение на рисунке 1 хоть и является непроективным, но может быть проанализировано без использования разрывных интерпретаций[8]. В общем случае можно утверждать следующее:

- Любая интерпретация, которую нельзя собрать без использования разрывных интерпретаций, является непроективной.
- Многие непроективные интерпретации можно собрать без использования разрывных интерпретаций.

Будем называть предложения, имеющие структуру, которая не может быть собрана без использования разрывных интерпретаций, *сверхнепроективными*. Примером сверхнепроективного предложения является фраза «*книгу я красную люблю*». Действительно, какая бы связь не была проведена первой, обязательно возникнет разрывная интерпретация. Обычно сверхнепроективные предложения встречаются в разговорной речи. Еще одним примером является классическая латинская поэзия, изобилующая такими предложениями [Ботвинник, Гладкий, 2009].

Однако существует широкий класс текстов, в которых встречаемость сверхнепроективных предложений очень низка. Кроме того, запрет на использование разрывных интерпретаций в процессе анализа значительно ускоряет работу анализатора, т. к. существенно уменьшает комбинаторику. Для того, чтобы контролировать появление разрывных интерпретаций, в Treevial были введены штрафы за пропуски. В управляющем файле с помощью специальной конструкции можно задавать величину штрафа, который взимается при возникновении всякой разрывной интерпретации. Важно понимать, что благодаря принципу инкрементальности штраф за разрыв остается «клеймом» на всех потомках разрывной интерпретации даже если в дальнейшем сам разрыв устраняется. Таким образом, все сверхнепроективные предложения всегда попадают в особую штрафную зону, что позволяет гибко настраивать анализатор при работе с корпусом.

---

[8] Разрывная интерпретация на рисунке 1 была приведена исключительно в иллюстративных целях.

4.6. Штрафы за некомпактность

В данном случае мы исходим из того, что в ситуации выбора из двух похожих по смыслу синтаксических структур более предпочтительной оказывается та, в которой зависимые в среднем расположены ближе к своим хозяевам. При проведении каждой новой связи взимается штраф, пропорциональный ее длине (расстоянию от хозяина до зависимого), умноженный на константу (число, которое задается с помощью конструкции языка Treevial). Понятно, что такая константа должна быть достаточно маленькой, т. к. фактор компактности не должен быть сильнее других. Хотя эта эвристика и не играет ключевой роли, она все равно кажется полезной. В ряде случаев с помощью нее удается, например, угадывать хозяев для предложных групп.

4.7. Механизм задания целевого шаблона

Для того, чтобы анализатор был способен обрабатывать не только полные предложения (в которых есть глагол в личной форме или предикатив), но и «кусочные» предложения (например, отдельные именные или предложные группы), синтаксис формального языка был расширен специальной конструкцией, позволяющей с помощью шаблонов и штрафных векторов описывать то, какие целевые структуры допустимы и какие из них предпочтительней (например: полное предложение на первом месте, далее именная группа, потом все остальные). В первую очередь, такая потребность была вызвана необходимостью анализировать заголовки стихотворений и литературных произведений.

4.8. Механизм учета лексической сочетаемости

Особняком стоит штрафной механизм, позволяющий учитывать при оценке синтаксических структур их конкретное лексическое наполнение, а именно сочетаемость лексических единиц, которые в процессе перебора были связаны некоторой синтаксической связью. Этот механизм был создан сравнительно недавно. Некоторые концептуальные вопросы, касающиеся его организации, до сих пор составляют предмет активных дискуссий авторов. Таким образом, можно сказать, что описание, приводимое ниже, носит эскизный характер.

Механизм учета лексической сочетаемости управляется не из основного управляющего файла, а обращается к лексической базе, в которой описаны различные ограничения на сочетаемость. За основу формального описания сочетаемости лексических единиц была взята модель, предложенная в [Перцов, Старостин, 1999]. В случае обнаружения несоответствия каких-то связей, проводимых в процессе анализа, модели сочетаемости, штраф соответствующих интерпретаций увеличивается.

В системе могут учитываться три типа ограничений на сочетаемость лексических единиц: морфо-синтаксические, лексические и семантические. К морфо-синтаксическим ограничениям на сочетаемость некоторого слова относятся ограничения на предложно-падежную форму зависимых слов (*агитация **[род]** [за+**вин**] [против+**род**]; блеснуть **[тв]** [перед+**тв**]; благодарный **[дат]** [за+**вин**]*).

Лексические ограничения на сочетаемость выражаются в том, что некоторые слова могут иметь в качестве зависимых только определенные лексемы, причем зачастую выделить общие признаки этих лексем оказывается невозможно. Например, в словосочетаниях *уменьшать / сбрасывать скорость, давление, громкость* глаголы *уменьшать / сбрасывать* имеют близкие значения, однако разную лексическую сочетаемость: можно уменьшить, но не сбросить длину, количество, сопротивление. Таким же образом описываются устойчивые выражения (глагол *обратить* сочетается с предлогом *на* только в выражении *обратить внимание на+**вин***).

Семантические ограничения отражены в лексической базе при помощи помет. Например, глагол *вернуться* имеет два актанта, отвечающих на вопросы *куда?* и *откуда?*, причем морфологически эти актанты могут быть выражены различными способами (*вернуться из Москвы в Петербург*, *вернуться издалека*, *вернуться домой*). В подобных случаях в лексической базе описываются не все возможные варианты выражения актантов, а указывается семантическая помета, которой актанты должны соответствовать (*вернуться **[НАПР] [ИСХ]***). Для некоторых лексических единиц также указано, как они трактуются в сочетании с различными наборами актантов (*издалека → **ИСХ***; *из **[род]** → **ИСХ***; *в **[пр]** → **ЛОК***; *в **[вин]** → **НАПР***). Для синтаксических интерпретаций, построенных на основе таких лексических единиц, строится их семантическая трактовка, которая затем при подчинении этих интерпретаций проверяется на соответствие семантическим ограничениям на сочетаемость подчиняющей интерпретации.

Рассмотрим два предложения: *Он познакомился с агитатором против истребления диких животных* и *Он выступал с друзьями против истребления диких животных*. Без использования механизма оценки сочетаемости анализатор не сможет отсеять неправильные гипотезы (в которых словосочетание *против истребления диких животных* зависит от глагола *познакомился* и существительного

друзьями соответственно). При использовании описываемого механизма анализатор предпочтет правильные гипотезы, поскольку только существительное *агитатор* (соответственно, глагол *выступать*) сочетается с предложно-падежной группой *против+**род***).

Рассмотрим еще один пример: *Встретил беженцев из Палестины* и *Вез сигареты из Москвы*. В лексической базе указана следующая информация о сочетаемости: *беженцы **[ИСХ]***, *вез **[вин] [ИСХ] [НАПР]***, *встретил **[вин]**, из **[род]*** → ***ИСХ***. Поскольку для предлога *из* в сочетании с родительным падежом указана семантическая трактовка *ИСХ*, для синтаксической интерпретации словосочетаний *из Москвы*, *из Палестины* будет построена семантическая трактовка *ИСХ*, которая в первом предложении сочетается только с существительным *беженцев*, а во втором — только с глаголом *вез*. Таким образом, приоритет получат правильные варианты анализа предложений (*беженцы из Палестины* и *вез из Москвы*).

Очевидно, что описания, которые можно заносить в лексическую базу, носят сильно упрощенный характер и ни в коей мере не претендуют на полноценное описание семантики лексических единиц. Однако, оказывается, что даже такие простые средства могут существенно улучшать качество синтаксического анализа. Важным свойством построенного механизма является то, что лексическая база может пополняться постепенно. При любой степени наполненности синтаксический анализ будет работать. Но чем больше информации будет внесено, тем точнее будет результат анализа. Следует отметить, что авторы далеки от мысли о ручном наполнении такой базы. В будущем планируется разработать методы автоматизированного извлечения сочетаемостных моделей из синтаксически-размеченных корпусов.

Заключение

В данной работе было описано то, как работает синтаксический анализатор Treevial, и изложены концепции, лежащие в его основе. В заключение следует отметить, что приведенная схема динамического ранжирования в принципе могла бы быть распространена на весь процесс анализа текста на естественном языке. «Поводы» для взимания штрафов имеют место на различных языковых уровнях. Приведем лишь некоторые из них:

- На нижних языковых уровнях
  - учет частотности употреблений словоформ
  - нечеткое сопоставление при распознавании звуков (символов) или текстов с ошибками
- На синтаксическом уровне
  - статистический учет сочетаемости слов
- На уровне семантики
  - использование семантических моделей, способных оценивать «осмысленность» синтаксической структуры

Авторы надеются, что впоследствии им удастся создать анализатор текстов, в рамках которого будут интегрированы все перечисленные функции. Архитектура анализатора Treevial уже сегодня ориентирована на расширение и подключение новых модулей.

Авторы горячо признательны Н. В. Перцову за многочисленные консультации и помощь в организации семинаров, посвященных анализатору Treevial. Мы благодарим всех участников семинара по вопросам автоматического лингвостиховедческого и морфосинтаксического анализа (ИМК МГУ / ЦТС ИРЯ РАН) за бурные дискуссии и обсуждения, а также С. А. Минора за ряд ценных советов. Особую благодарность авторы выражают И. А. Пильщикову — за помощь в подготовке окончательного текста данной статьи.

Литература

1. *[Ботвинник, Гладкий, 2009]* — Гладкий А. В., Ботвинник Н. М. «Переплетение слов» в русской и латинской поэзии // «Слово — чистое веселье...»: Сборник статей в честь Александра Борисовича Пеньковского. М.: Языки слав. культуры, 2009. С. 299–310.
2. *[Гладкий, 1985]* — Гладкий А. В. Синтаксические структуры естественного языка в автоматизированных системах общения // М.: Наука, 1985.
3. *[Гладкий, Мельчук, 1969]* — Гладкий А. В., Мельчук И. А. Элементы математической лингвистики // М.: Наука, 1969.
4. *[Мальковский, Старостин, 2006]* — Мальковский М. Г., Старостин А. С. Модель синтаксиса в системе морфосинтаксического анализа «TREETON» // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2006». М.: изд-во РГГУ, 2006. С. 481–492.
5. *[Мальковский, Старостин, 2007]* — Мальковский М. Г., Старостин А. С. Алгоритм синтаксического анализа, используемый в системе морфо-синтаксического анализа «TREETON» // Труды международной конференции Диалог 2007. М.: изд-во РГГУ, 2007. С. 516–524.
6. *[Перцов, Старостин, 1999]* — Перцов Н. В., Старостин С. А. О синтаксическом процессоре, работающем на ограниченном объеме лингвистических средств // Труды международной конференции Диалог 1999. Таруса: 1999. Т. 2. С. 224–230.
7. *[Тестелец, 2001]* — Тестелец Я. Г. Введение в общий синтаксис // М.: РГГУ, 2001.
8. *[Хомский, 1962]* — Хомский Н. Синтаксические структуры // Новое в линвистике. Вып. II. М., 1962. С. 412–526.
9. *[Cocke, Schwartz 1970]* — John Cocke and Jacob T. Schwartz Programming languages and their compilers: Preliminary notes // Technical report Courant Institute of Mathematical Sciences, New York University, 1970.
10. *[Dekhtyar, Dikovsky, 2008]* — Michael I. Dekhtyar, Alexander Ja. Dikovsky. Generalized Categorial Dependency Grammars // Pillars of Computer Science, 2008. P.230–255
11. *[Johnson, 1998]* — Johnson M. PCFG models of linguistic tree representations // Computational Linguistics, 1998. №24(4). P. 617–636.
12. *[Kasami, 1965]* — Kasami T. An efficient recognition and syntax-analysis algorithm for context-free languages // Scientific report Afcrl-65-758, Air Force Cambridge Research Lab, Bedford, MA., 1965
13. *[Mel'cuk , 1988]* — Mel'cuk I. A. Dependency Syntax: Theory and Practice // SUNY Series in Linguistics. Albany: State University of New York Press, 1988
14. *[Pollar, Sag, 1994]* — Carl Pollard, Ivan A. Sag. Head-Driven Phrase Structure Grammar // Chicago: University of Chicago Press, 1994.
15. *[Schneider, 1998]* — Garold Schneider. A Linguistic Comparison of Constituency, Dependency and Link Grammar // http://www.ifi.unizh.ch/cl/study/lizarbeiten/lizgerold.pdf, 2007
16. *[Sleator & Temperley, 1991]* — Sleator D., Temperley D. Parsing English with a Link Grammar // Carnegie Mellon University Computer Science technical report CMU-CS-91-196, 1991.
17. *[Tesnière 1959]* — Tesnière L. Éléments de syntaxe structurale // Paris: 1959.
18. *[Younger, 1967]* — Daniel H. Younger. Recognition and parsing of context-free languages in time n3 // Information and Control, 1967. № 10 (2). P. 189–208.

Звуковая реальность словоизменительных аффиксов (по данным звукового корпуса русского языка)[1]

Phonetic realization of russian inflections in the ord speech corpus of everyday communication

Степанова С. Б. (stsvet_2002@mail.ru),
Асиновский А. С. (a.s.asinovsky@gmail.com),
Рыко А. И.(aryko@mail.ru),
Шерстинова Т. Ю. (sherstinova@gmail.com)

Факультет филологии СПбГУ

На материале записи спонтанных монологов из Звукового корпуса русского языка проведен анализ словоизменительных аффиксов. Выявлены наиболее часто встречающиеся грамматические значения, передаваемые флексиями, их возможные фонетические варианты и случаи недореализации флексий.

Общеизвестным является факт, что русский язык относится к языкам флективного строя, то есть грамматическое значение словоформы (род, число, падеж, лицо) выражается в русском языке через окончание (аффикс), которое служит для указания на синтаксические отношения, связи данного слова с другими словами во фразе. При этом практически все русские флексии являются омонимичным: одно и то же звучание может выражать различные грамматические значения: так, флексия [oj] может иметь значение «прилагательное мужского рода, единственного числа, именительный падеж — *большой*» или «существительное женского рода, единственного числа, творительный падеж — *с добротой*» и др. Думается, при создании систем автоматического распознавания речи небесполезными могли бы быть сведения, какие грамматические значения одной и той же звуковой манифестации являются наиболее частотными.

С другой стороны, флексии, находясь в артикуляторно слабой позиции (заударный слог, конец слова, конец фразы), в наибольшей степени подвержены качественной редукции и отличаются максимальной вариативностью реализаций. Анализ реального звучания словоформ и, в частности, словоизменительных аффиксов на материале представительного корпуса спонтанной речи — одно из направлений исследований, выполняемых на материале ЗКРЯ, первые результаты такого исследования представлены в данном докладе.

Для проведения **корпусного** исследования реализаций словоизменительных аффиксов в спонтанной русской речи, в качестве первого шага, для нескольких файлов из Звукового корпуса русского языка (ЗКРЯ) было осуществлено аннотирование по методу сплошной выборки: вычленялись и транскрибировались **все** встречающиеся в материале флексии существительных, прилагательных, финитных форм глаголов, причастий, порядковых числительных, изменяемых разрядов местоимений (притяжательных, указательных и т. д.).

Необходимо сразу оговориться, что в некоторых случаях, мы отходили от классического понимания флексии и выделяли «финаль слова» — так называемый *морфный комплекс*, завершающий словоформу. Это относится к существительным на *-ие* (напр., *поколен{ij-e}, настроен{ij-e}* и т. п.): здесь мы нарушаем традиционный принцип выделения флексии и рассматриваем финальную последовательность *-ие* как единый элемент. Мы сочли возможным «отдать» флексии гласную основы, так как:

- во-первых, финаль в подавляющем большинстве случаев реализовалась как достаточно краткий, однородный гласный, делить который на основу и флексию мы сочли нецелесообразным;

[1] Исследование выполнено при поддержке гранта РГНФ «Разработка информационной среды для мониторинга устной русской речи» (09-04-12115в).

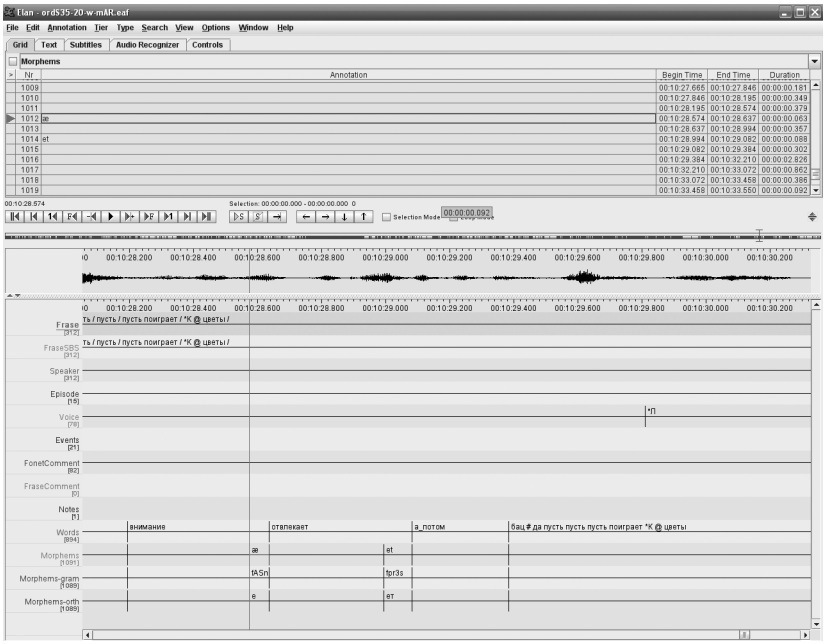


Рис. 1. Аннотирование на уровне Morphems-gram и Morphems-orth в программе ELAN(часть фразы из файла ordS35-20: ага / это он сейчас внимание отвлекает / а потом бац / # да пусть / пусть / пусть поиграет / *К @ цветы /)

- во-вторых, считаем маловероятным, что эта финаль может на фонетическом уровне реализоваться как-то иначе, чем, например, флексия в прилагательных (*синие*).

Процедура выделения флексии методом слухового анализа, поддержанного возможностью наблюдать реализацию звучания на спектрограмме и осциллограмме, а также транскрибирование слухового впечатления от флексии проводились в программе Praat. Транскрипция осуществлялась в символах МФА. В случае если морфема не была реализована или «скрылась» в предыдущем согласном, выделялся короткий символический отрезок в конце словоформы, который обозначался знаком [-] (отсутствие реализации, полная редукция).

Затем файл Praat конвертировался в файл программы ELAN и проводилось аннотирование флексий на уровнях Morphems-gram (грамматическое значение морфемы) и Morphems-orth (орфографическое представление морфемы) — см. рис. 1.[2]

Для аннотирования грамматического значения словоизменительных флексий мы использовали следующую систему обозначений[3]. Первым символом в каждом обозначении является «f» — указание на принадлежность элемента к классу флексий.

1. Для имен — указание на падеж и число:
fNS — именительный ед.ч., fGS — родительный ед.ч., fG2S — «второй родительный» (партитивный) ед.ч., fDS — дательный ед.ч., fAS — винительный ед.ч., fASi — винительный ед.ч. (для неодушевленных существительных мужского и среднего рода и согласованных с ними прилагательных), fIS — творительный ед.ч., fLS — предложный ед.ч., fL2S — местный ед.ч. (*в лесу*), fNP — именительный мн.ч., fGP — родительный мн.ч., fDP — дательный мн.ч., fAP — винительный мн.ч., fAPi — винительный мн.ч. (для неодушевленных существительных и согласованных с ними прилагательных), fIP — творительный мн.ч.,fLP — предложный мн.ч.

1.2. Для существительных после указания на падеж и число отмечен тип склонения:
fNS1 — 1-е склонение (*сестра, земля*), fNS2 — 2-е склонение (*кот, конь; село, поле*), fNS3 — 3-е склонение (*лошадь*), fNS4 — существительные 2-го склонения с исходном основы на *–j*, которые, по нашим данным, образуют особый тип, в котором невозможно отделить окончание от финали основы *(санаторий)*.

1.3. Для прилагательных (а также местоимений-прилагательных, порядковых числительных) после указания на падеж и число отмечаем род или принадлежность к подпарадигме мн.ч. :
fNSm — мужской, fNSf — женский, fNSn — средний, fNPp — подпарадигма мн.ч.

2. Глаголы

2.1. Подсистема настоящего/будущего времени.

Отмечаем принадлежность к этой подсистеме (pr), лицо (1, 2, 3) и число (s, p):
fpr1s (*хожу),* fpr2s (*ходишь),* fpr3s (*ходит),* fpr1p (*ходим),* fpr2p (*ходите),* fpr3p (*ходят)*

---

[2] О системе аннотирования на других уровнях см.(Шерстинова, Рыко, Степанова 2009)

[3] В силу специфики данного конкретного исследования нулевая флексия не выделялась и не аннотировалась.

2.2. Подсистема прошедшего времени.

Отмечаем принадлежность к этой подсистеме (p) и к роду (множественному числу):

fpf (*ходила),* fpn (*ходило),* fpp (*ходили)*

2.3. Краткие пассивные причастия прошедшего времени

fpartf (*написана),* fpartn (*написано),* fpartp (*написаны)*

Всего к настоящему времени таким образом проаннотировано 880 морфем из 4 файлов корпуса ОРД .

В табл.1 представлены результаты автоматической обработки морфемных уровней аннотирования. Приведены флексии, встретившиеся в проанализированном материале не меньше 10 раз.

Таблица 1. Частота встречаемости морфем с различными грамматическими значениями

КОД грамматического значения	Абсолютное количество	%
fNS1	72	8,18
fNSm	71	8,07
fpf	52	5,91
fpr2s	46	5,23
fpr3s	43	4,89
fAS1	41	4,66
fNSf	35	3,98
fpp	35	3,98
fNP2	33	3,75
fGS1	26	2,95
fpr1s	26	2,95
fGS2	25	2,84
fNPp	24	2,73
fpr3p	23	2,61
fNSn	22	2,50
fpr1p	20	2,27
fISm	19	2,16
fASf	18	2,05
fASim	18	2,05
fLS2	16	1,82
fLS1	15	1,70
fNS4	14	1,59
fAPi2	11	1,25
fNS2	11	1,25
fNP1	10	1,14

Таблица 2. Частота встречаемости различного типа морфем, представленных в одинаковой орфографической форме

Орфографическое представление	Абсолютное количество	%
а	172	19,55
и	92	10,45
ой	63	7,16
у	62	7,05
е	37	4,20
ешь	36	4,09
ый	34	3,86
о	30	3,41
ая	29	3,30
ы	29	3,30
ие	22	2,50
ое	22	2,50
ю	22	2,50
ет	21	2,39
ит	18	2,05
им	17	1,93
ом	16	1,82
ые	13	1,48
ей	12	1,36
ем	12	1,36
ий	12	1,36
ого	9	1,02
ют	9	1,02
ишь	7	0,80
ут	7	0,80
ую	7	0,80
ия	6	0,68
ым	6	0,68
я	6	0,68
ят	6	0,68
ете	5	0,57
ём	4	0,45
ёт	4	0,45
ов	4	0,45
ых	4	0,45
ами	3	0,34
ёшь	3	0,34
ими	3	0,34
ому	3	0,34
ах	2	0,23
ев	2	0,23
ии	2	0,23
ях	2	0,23
ам	1	0,11
ат	1	0,11
его æ	1	0,11
ием	1	0,11
ию	1	0,11
Всего:	880	

Сопоставление этих таблиц позволяет сделать некоторые предварительные наблюдения над функционированием словоизменительных аффиксов в русской спонтанной речи и наметить пути их дальнейшего исследования.

1. Типы и реализации флексий, представленных орфографическим {а}

Как видно из табл. 2, самой частотной орфографической манифестацией флексии на нашем материале является {а} — почти 20 % от всех словоизменительных финалей. Рассмотрим, какие грамматические значения она чаще всего выражает и каким образом реализуется.

В табл.3 видно, что самое частотное значение {а} — Им. п. имен существительных ед. ч. 1-ого склонения (*мама, папа*). Чуть реже встречается глагольное окончание (ж. род, прош. вр.: *была, делала*), значительно реже — Род. п. ед. ч. существительных 2-ого склонения (*друга*). На долю остальных 9 значений {а} приходится около 13 % её употреблений в проанализированном материале.

Таблица 3. Количество различных грамматических значений морфемы, представленной {а}

Грамматическое значение	Абсолютное количество	%
fNS1	68	39,535
fpf	52	30,233
fGS2	25	14,535
fNP2	8	4,6512
fAS2	6	3,4884
fNSf	6	3,4884
fASi	2	1,1628
fAP2	1	0,5814
fAPi7	1	0,5814
fASf	1	0,5814
fFS2	1	0,5814
fpartf	1	0,5814
	172	100

Реализация {а} характеризуется максимальным разнообразием представленных аллофонов. Впрочем, об этом, как об особой изменчивости именно фонемы /а/ в зависимости от позиционных и комбинаторных условий, неоднократно писали фонетисты (см., напр.: Бондарко 1998; Кузнецов 1997; и др.). Обратим внимание на самые частотные аллофоны /а/.

Необходимо сразу оговориться, что≅ в нашей транскрипции имеется в виду под тем или иным знаком МФА.

При транскрибировании самыми частыми обозначениями являлись следующие знаки для аллофонов /а/:

[а] — гласный среднего ряда нижнего подъема, более или менее соответствующий изолированному произнесению /а/ в русском языке. Этот знак в МФА используется для обозначения переднего гласного, однако мы выбрали именно его для гласного среднего ряда (а не[ɐ]) из соображений максимального удобства при транскрибировании самого частотного русского аллофона фонемы /а/.

[ɑ] — гласный заднего ряда, нижнего подъема,

[ʌ] — гласный заднего ряда средне-нижнего подъема, отличающийся от [ɑ] не столько подъёмом, сколько краткостью, и именно из-за краткости реализующийся как менее продвинутый назад гласный,

[æ] — гласный переднего ряда средне-нижнего подъема.

Остальные знаки совпадают по значениям с теми описаниями, которые имеются в литературе и в Интернете относительно МФА.

В табл. 4 представлено процентное соотношение различных аллофонов, реализующих финальное {а}.

Таблица 4. Аллофоны, реализующие финальное {а}

а	45	ɵ	2
ə	33	aj	1
-	17	e	1
ʌ	16	u	1
ɑ	13	ɑ̯	1
æ	11	ɐ	1
ɜ	6	ã	1
ə̯	5	əj	1
ɞ	4	ɻ	1
ɨ	3	ɔ̃	1
ə̯	3	oe	1
ɔ	3	æa	1
			172

Видно, что чаще всего морфема {а} реализуется в нашем материале как [а] или как [ə] (45 и 33 употребления соответственно).

Таблица 5 показывает, каким образом конечное [а] реализуется в разных флексиях {а}.

Таблица 5. От реализации — к грамматическому значению

ɑ		ʌ		æ	
fpf	8	fpf	11	fNS1	7
fGS2	2	fNS1	5	fNSf	6
fNS1	2	fNS2	1	fGS2	4
fAS2	1	fNSm	1	fGS4	2
fNS2	1	fpartn	1	fAPi2	1
				fAS4	1
				fLS2	1
				fNP1	1
				fNPp	1
				fNSn	1
				fpf	1

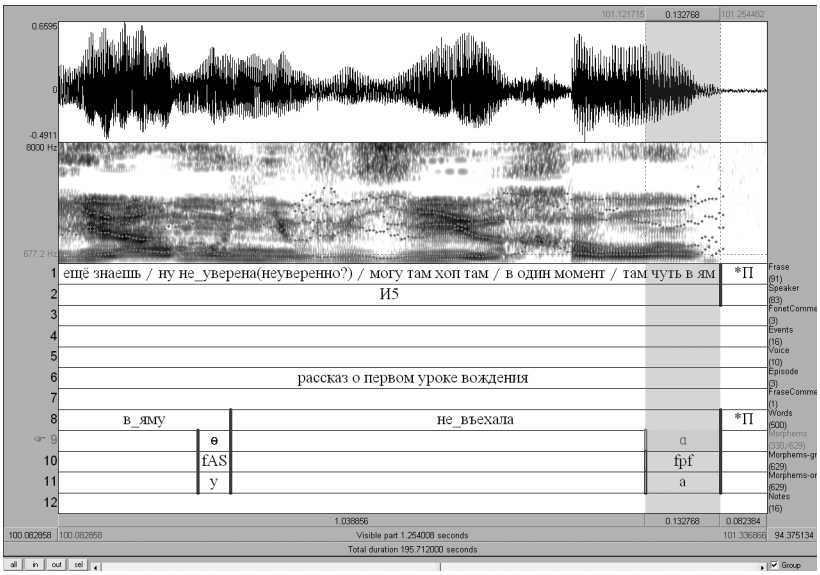


Рис. 2. Осциллограмма, спектрограмма, многоуровневая разметка словоформы ***не въехала*** из фразы ***…там чуть в яму не въехала*** // *П

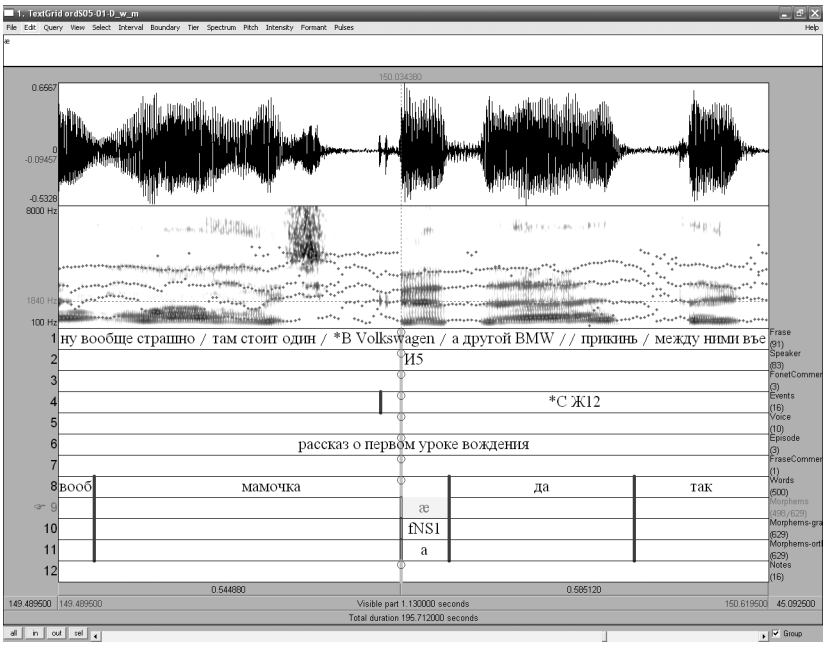


Рис. 3. Осциллограмма, спектрограмма, многоуровневая разметка словоформы ***мамочка*** из фразы ***вообще мамочка! да / так***

Обращает внимание на себя тот факт, что как задний гласный [ɑ] или [ʌ] реализуется чаще всего морфема со значением «прошедшее время глагола, ж. р.», а аллофон [æ] практически не встречается в морфеме с этим значением. Причем акустическое и перцептивное различие между [ɑ] и [æ] оказалось весьма существенным. На рис. 2 и 3 можно видеть осциллографическое и спектральное представления аллофонов гласного /а/ из флексий в словах *въехала* и *мамочка.*

Различие в спектрах очевидно: средние значения формант4 [ɑ] из *въехала* F1 — 700 гц, F2 — 1200 гц, [æ] из *мамочка* — F1 — 600 гц, F2 — 1700 гц. В результате при изолированном прослушивании этих гласных они воспринимаются как два абсолютно непохожих звука. Однако это различие не связано с противопоставлением разных морфем {а}, оно имеет чисто фонетическую природу. Задний характер гласного во флексиях глаголов прошедшего времени вызван исключительно соседством с сильно велярезованным согласным /l/ . Неожиданный после заднеязычного согласного передний характер гласного [æ] во флексии слова *мамочка* можно объяснить мягким [č] перед [k] и последующим, произнесенным без паузы, *да* с переднеязычным согласным.

[4] Точечные линии на рисунках — значения формант.

О других аллофонах гласного /a/ никак нельзя сказать, что они могут быть привязаны к какой-то определенной грамматической форме — см. табл. 6 (еще раз подчеркнем, что и со значением прошедшего времени аллофоны [ɑ] и [ʌ] соотносятся лишь через суффикс [l]).

Таблица 6. От реализации — к грамматическому значению

[a]		æ		ə		полная редукция	
fNS1	13	fNS1	7	fNS1	14	fNS1	9
fpf	13	fNSf	6	fNSm	9	fpf	5
fNSf	8	fGS2	4	fpf	9	fAS1	4
fGS2	7	fGS4	2	fGS2	6	fNSm	4
fNP2	3	fAPi2	1	fpp	5	fpr1s	4
fNS2	3	fAS4	1	fGS1	4	fGS1	3
fNSn	3	fLS2	1	fASn	3	fNP2	2
fpn	3	fNP1	1	fGSf	3	fpn	2
fAS2	2	fNPp	1	fNSf	3	fpp	2
fAP2	1	fNSn	1	fNSn	3	fpr2s	2
fASf	1	fpf	1	fpn	3	fAPi1	1
fASi2	1			fASf	2	fASf	1
fASim	1			fASi2	2	fASi2	1
fDS2	1			fLSf	2	fGS3	1
fpartn	1			fNP2	2	fLS1	1
				fpartn	2	fLS2	1
				fAPi2	1	fNS2	1
				fAS1	1	fNSn	1
				fAS2	1	fpartf	1
				fGSm	1	fpr3s	1
				fIS1	1		
				fISf	1		
				fLS1	1		
				fNPp	1		
				fNS2	1		
				fpr3p	1		

Во всех случаях максимально частотное появление аллофонов /a/ — в словоформах существительных ж. рода Им. п. ед. ч. (fNS1), что связано с общей частотностью этой формы в речи наших информантов.

2. Типы и реализации флексий, представленных орфографическими {и} и {е}

Таблица 7 показывает, что чаще всего одиночная гласная {е} представляет окончание пр. п. существительных ед. ч. 1-ого или 2-ого склонения (30 из 36 встретившихся в материале {е}). Одиночное {и} чаще всего встречается в морфемах, имеющих значение «прошедшее время глагола, мн. ч.», «им. п. существительных, мн. ч.» и «род. п. существительных, ед. ч., 1-ое склонение».

Таблица 7. От орфографии — к грамматическому значению

е		и	
fLS2	16	fpp	35
fLS1	14	fNP2	17
fDS1	3	fGS1	15
fGS1	1	fNP1	6
fNS2	1	fNPp	5
fASɨ	1	fAPi7	3
		fGS3	3
		fLS3	3
		fAPɪ	1
		fAPi3	1
		fAPip	1
		fASɪ	1
		fLS1	1

Вопрос о реализации этих морфем в безударном положении тесно связан с проблемой э≅кающего или и≅кающего произношения в современной русской речи, которая активно обсуждалась в русистике последних десятилетий. Можно сказать, что точку в дискуссии поставили исследования, проведенные под руководством Л. А. Вербицкой, в которых была проанализирована речь 150 ленинградцев. Экспериментальные данные, полученные в ходе инструментального анализа убедительно показали, что ведущим в конце XX века стал и≅кающий вариант нормы (Вербицкая 1976).

Нисколько не оспаривая этого факта (в нашем распоряжении пока лишь описание 880 флективных морфем из речи 4-х дикторов), мы вынуждены признать, что наш материал предоставляет и другие варианты реализации гласных на месте как орфографического {е}, так и орфографического {и} (см. табл. 8).

Таблица 8. Реализации гласных на месте орфографических {е} и {и}

Реализации Е	Абс. кол-во	%	Реализации Е	Абс. кол-во	%
e	13	35,1	i	40	43,5
i	6	16,2	e	15	16,3
ɛ	4	10,8	ɪ	10	10,8
–	3	8,1	ə	7	7,6
y	2	5,4	i̥	5	5,4
ɪ	1	2,7	—	4	4,3
i̥	1	2,7	i̯	2	2,2
ĩ	1	2,7	i͡	2	2,2
iæ	1	2,7	ea	1	1,1
jo	1	2,7	i:	1	1,1
ə	1	2,7	ш	1	1,1
з	1	2,7	y	1	1,1
ɵ	1	2,7	e	1	1,1
æ	1	2,7	oe	1	1,1
	37	100	æ	1	1,1
				92	100

Причем {е} как [е] наблюдалась не только в ударном слоге (что естественно), но и в безударных окончаниях.

Реализация {и} как [і] преобладает во флексиях (43,5 %), однако и здесь возможны реализации более открытого гласного — [е] (16 %) и др. Это наводит на мысль, что закрытые передние гласные вследствие вялой артикуляции в конечной безударной позиции отодвигаются чуть назад и чуть вниз, что и дает восприятие гласного как *е*-образного. Несомненно, подготовленный к настоящему времени материал не дает возможности сказать, насколько это явление распространено в современной спонтанной речи. Возможно, это характерно лишь для некоторых наших информантов.

3. К постановке вопроса о реализации финалей, представленных некоторыми орфографическими двухбуквенными сочетаниями

Речь в данном случае идет главным образом о тех явлениях, о которых писала Л. В. Бондарко: «в словах *тихий, серый* нет никаких следов конечного /j/, вообще неустойчивость среднеязычного сонанта проявляется в заударных комплексах, где он оказывается между гласными. В этих случаях единственным следом сонанта оказывается передний характер гласного, который следует за ним» (Бондарко 1998: 266). Именно этот тезис можно подтвердить, наблюдая материал, обработанный к настоящему времени.

Из 46 финалей {ый}/{ий} в различных формах /j/ был реализован лишь в одном случае (см. табл. 9)!

Иначе обстоит дело с окончанием {ой}. В табл. 10 можно видеть, что чаще всего эта финаль встречается в окончаниях имен прилагательных (Им. п., м. род, ед. ч.). Как известно, это окончание всегда ударно. Думается, именно поэтому почти в 30 % случаев конечное буквосочетание {ой} реализовано как [oj].

Что касается заударных морфных комплексов, где среднеязычный сонант оказывается между гласными, то, как показывает наш материал (табл.11), исчезает не только этот сонант, но и само сочетание гласных в подавляющем большинстве случаев стягивается до одного однородного аллофона (ср.: Овчаренко 1988:9).

Таблица 9. От орфографии — к реализации

ий		ый		ой	
i	8	ɨ	11	oj	20
ɨ̞	3	ə	7	ə	8
ɪ	1	–	4	ɵ	7
Всего:	12	ɨ̃	3	o	6
		ɨ̞	2	ɔ	4
		a	1	ɨ	3
		i	1	əj	3
		ɪ	1	ij	2
		ij	1	oe	2
		y	1	e	1
		ʌ	1	ɪ	1
		з	1	oi	1
		Всего:	34	õ	1
				u	1
				ʊj	1
				ɞ	1
				ø	1
				Всего:	63

Таблица 10. От орфографии — к грамматическому значению

ой	
fNSm	36
fASim	7
fGSf	6
fLSf	5
fIS1	4
fISf	3
fDSf	2

Из 99 финалей только в 28 реализовано бифонемное сочетание. Думается, они приходятся на ударные окончания, что предстоит еще проверить.

В целом можно сказать, что корпусный подход к накоплению, систематизации и — что особенно важно — к анализу материала русской звучащей речи позволяет проверить многие сделанные ранее наблюдения фонетистов, подтвердить или опровергнуть многие научные гипотезы, ответить на многие поставленные ранее вопросы и, может быть, поставить новые, на которые предстоит искать ответам новым поколениям лингвистов.

Таблица 11. От орфографии — к реализации

ая		ие		ия		ое		ую		ые	
æ	6	e	10	æ	2	oə	5	u	2	ɨ	3
a	3	i	5	i	1	ə	4	uj	1	ɨə	3
aə	3	ɪ	3	ɨə	1	a	2	uy	1	ɨ-ə	3
aз	2	æ	2	iæ	1	з	2	ʉ	1	ɨe	1
ə	2	e̥	1	æ̃	1	–	1	ʊ	1	ɨʒ	1
æe	2	iə	1	Всего	6	o	1	ʊə	1	ə	1
ae	1	Всего	22			oə	1	Всего	7	ɛ	1
e	1					oi	1			Всего	13
iæ	1					oj	1				
ɑj	1					oy	1				
a-ə	1					oз	1				
ḁ	1					oæ	1				
əə	1					ɛ	1				
ə̃	1					Всего	22				
з	1										
ɵ	1										
æj	1										
Всего	29										

Литература

1. *Бондарко Л. В.* Фонетика современного русского языка. СПб., 1998.
2. *Вербицкая Л. А.* Русская орфоэпия: К проблеме экспериментально-фонетического исследования особенностей современной произносительной нормы. Л., 1976.
3. *Кузнецов В. И.* Вокализм связной речи. СПб., 1997.
4. *Овчаренко Е. Б.* Реализация и восприятие заударных морфных комплексов и функциональная нагрузка морфем. // Автореф. канд. дисс. Л., 1988.
5. *Шерстинова Т. Ю., Рыко А. И., Степанова С. Б.* Система аннотирования в звуковом корпусе русского языка «Один речевой день». // Формальные методы анализа русской речи. Материалы секции 18-ой Международной конференции. СПбГУ, 2009. С. 66–75.

Словообразовательная разметка Национального Корпуса русского языка: задачи и методы

Word-formation annotation of the Russian National Corpus: aims and methods

Тагабилева М. Г. (geratagabileva@gmail.com),
Березуцкая Ю. Н. (onthehay@mail.ru)
Московский Государственный Университет

1. Введение

Практика создания современных корпусов предполагает в первую очередь разметку данных на уровне слова (например, разметку лемм, частей речи, грамматических признаков и т. д.), а также единиц, более крупных, чем слово (например, разметку синтаксических групп, коммуникативного членения предложения, приведение сведений о тексте в целом и т. д.). Нужна ли в корпусе информация о двусторонних единицах, меньших чем слово, таких, как корни, приставки и суффиксы?

В данной статье мы хотим обрисовать перспективы корпусно-ориентированного подхода к изучению русского словообразования, а также показать, какие практические возможности может предоставить словообразовательная разметка пользователям корпуса.[1]

Существующие описания русского словообразования включают в себя более или менее полные перечни словообразовательных моделей (например, *под — √ — ени — е: вед, ключ, нес, нош, тверждд, чин* [Кузнецова, Ефремова 1986], ср. также [Кубрякова 1965, Townsend 1968, Шанский 1968, Земская 1973, 1992, Развитие.. 1975., Улуханов 1977, 1996, РГ 1980, Шанский, Тихонов 1981, Ефремова 2000]) и списки словообразовательных гнезд (например, *смелый, смело, смелость, смельчак, смелеть, осмелеть, осмелиться, посмелеть*; см. словообразовательные словари [Тихонов 1985, Потиха 1961, Тихонов 1978, Wolkonsky, Poltoratzky 1969] и словарь морфем [Кузнецова, Ефремова 1986]). Однако есть проблема, которая, как нам кажется, все еще не нашла удовлетворительного решения — она касается продуктивности словообразовательных моделей. В настоящее время под этим термином понимается прежде всего продуктивность в словаре. Продуктивность оценивается относительно всего словарного состава языка (например, непродуктивна модель с суффиксом *б(а)*, ср. *судьба*) или определенного лексического класса (например, модель *с суффиксом ец* непродуктивна для класса слов с адъективным корнем, ср. *глупец*, в то время как модель с суффиксом *ин(а)* продуктивна для класса названий животных, ср. *оленина*). Продуктивность может также оцениваться в диахроническом ключе, но опять-таки сквозь призму словаря (ср. появление суффикса *ирова(ть)* в XVII в. [Изменения... 1964: *44*], вовлечение в словообразовательные модели заимствованных и новых слов).

Вместе с тем продуктивность можно понимать и как вероятность реализации словообразовательной модели в тексте. В самом деле, легко представить себе иностранца, которому непонятен выбор — в конкретном контексте — между диминутивом и недиминутивной формой, между разными моделями образования отглагольного имени и т. д. Не следует также забывать, что для носителя языка словообразование — это живая деятельность, проявляющаяся в речи в виде окказионализмов, языковых игр и т. д.; ср. показательное название книги Е. А. Земской «Словообразование как деятельность»

[1] Данная работа выполнена в рамках Программы фундаментальных исследований ОИФН РАН «Текст во взаимодействии с социокультурной средой: уровни историко-литературной и лингвистической интерпретации (2009–2011гг.)»

[Земская 1992]. До сих пор продуктивность словообразования в тексте и речи изучалась в основном в стилистическом аспекте [Виноградова 1984 и др.]. Однако, как кажется, любопытно было бы проанализировать, как реализация словообразовательных моделей в тексте связана с реализацией других конструкций; как одни словообразовательные модели сочетаются с другими; как отличается поведение глагольных и именных корней; каким образом однокоренные слова задействованы в установлении кореференции; какова частотность той или иной модели в корпусе в целом или в том или ином жанре, ср. [Пазельская 2009]; также небезынтересно было бы проследить микро-изменения в процессах словообразования (например, какова скорость вовлечения в словообразование новых слов и др.).

Все эти возможности может предоставить словообразовательная разметка корпуса, выполненная с привлечением электронного словообразовательного словаря и снабженная поисковой системой. Данная статья представляет собой проспект проекта, нацеленного на создание полноценного словообразовательного модуля в Национальном корпусе русского языка (http://ruscorpora.ru). Излагаются задачи первого этапа — составления словообразовательной базы данных, ориентированной на разметку корпуса. Поскольку база данных представляет по сути словарь, но реализованный в электронном виде, у нас есть возможность совместить два формата — традиционный словарь морфем и традиционный словообразовательный словарь, то есть можно будет с его помощью выяснить морфемное членение интересующего слова, найти все слова с конкретным корнем (словообразовательное гнездо) или же списки слов с тем или иным аффиксом или сочетанием морфем (словообразовательную модель). Основной акцент при разметке базы данных делается на кодировании плана выражения словообразовательных единиц — их алломорфов, чередований и порядка следования.

Далее в статье мы проиллюстрируем предусмотренные поисковые возможности (раздел 2); коснемся общетеоретических проблем (раздел 3), очертим план работы и возможные подходы к автоматизации разметки (разделы 4 и 5).

2. Предусмотренные поисковые возможности

Поскольку словообразовательная разметка НКРЯ — первый проект подобного рода в практике аннотации лингвистически ориентированных корпусов, то одной из первоочередных задач стало составление списка возможных поисковых запросов к корпусу со словообразовательной разметкой. Именно от круга задач, которые может поставить перед корпусом пользователь, зависят формат и степень подробности самой разметки.

Представляется, что самым распространенным поисковым запросом должен стать поиск слов, содержащих конкретную морфему (возможно, конкретный алломорф какой-либо морфемы) или некоторое определенное сочетание морфем. Это дало бы пользователю возможность исследовать лексемы, образованные по конкретной словообразовательной модели, анализировать сочетаемость морфем и частотность тех или иных сочетаний, влияние той или иной морфемы на значение содержащего ее слова и особенности его употребления в тексте, свойства слов, принадлежащих к одному словообразовательному гнезду, считать частотность однокоренных слов разных частей речи и решать еще довольно широкий круг теоретических и практических вопросов. В связи с тем, что принадлежность того или иного алломорфа к конкретной морфеме — факт зачастую не очевидный, необходимо предоставить пользователю доступ к списку морфем и их алломорфов: таким образом, можно будет не задавать искомый морф(ему) вручную, а выбирать из представленного списка. К сожалению, словарь морфем А. И. Кузнецовой и Т. Ф. Ефремовой дает список только корневых и префиксальных алломорфов, сведенных в морфемы, но не дает подобных списков для суффиксов, и сведение суффиксальных алломорфов в морфемы — одна из самых сложных теоретических задач, которые нам еще предстоит решить.

Кроме того, необходимо предусмотреть возможность поиска не только по конкретным значениям параметров, но и по наличию/отсутствию помет того или иного типа, то есть поиск композитов, поиск слов, содержащих один или более аффиксов (например, вполне возможным представляется запрос «найти все слова с двумя приставками»), поиск слов со связанными корнями. Естественно, нужно предоставить пользователю корпуса возможность комбинировать поиск обоих типов (это позволит искать, например, все префиксальные (суффиксальные) производные от конкретного корня). Также должна существовать опция комбинирования поиска по словообразовательной разметке с поиском по другим (семантическим и грамматическим) параметрам, предусмотренным разметкой НКРЯ.

3. Общетеоретические проблемы

Русский язык обладает чрезвычайно обширным инвентарем словообразовательных средств и разветвленной системой правил взаимодействия морфологии и фонологии. Именно поэтому в самом начале нашей работы одной из главных задач стало выявление общетеоретических проблем, которые могут встать перед разметчиком или разработчи-

ком алгоритма автоматической разметки словника, а также выработка системы принципиальных последовательных решений. Поскольку основная цель проекта — в первую очередь практическая (а именно, словообразовательная разметка Национального корпуса русского языка, а не фундаментальное описание системы русского словообразования как таковое), то и общее направление поиска решений было ориентировано на максимальную формализацию и упрощение процесса разметки.

Основные теоретические проблемы, осложняющие процесс морфемного анализа, подробно описаны в [Кузнецова, Ефремова 1986: *3–9*]. К ним относятся в первую очередь проблемы семантики (степень опоры на семантику аффиксов при словообразовательном анализе), омонимия (омоморфия, проблема омонимичных аффиксов), проблема эквивалентных решений («Одно и то же слово в силу многообразия структурно-семантических ассоциаций его с другими словами языка можно соотносить с несколькими мотивирующими словами (основами). Это неизменно приводит к появлению параллельных синхронных разночтений деривационных структур… и морфемного состава слова, особенно его посткорневой части…» [Кузнецова, Ефремова 1986: 6]).

Первым вопросом стал инвентарь используемых в разметке терминов — названий классов морфем. Наряду с выделяемыми в классических теориях префиксом, корнем, суффиксом, и интерфиксом («соединительной гласной»), в современных исследованиях присутствуют и так называемые аффиксоиды (префиксоид и суффиксоид — морфемы, способные сочетаться и с бесспорными корнями, и с бесспорными аффиксами). Вопрос об их существовании, в том числе и в русском языке, — один из спорных и активно обсуждаемых морфологами вопросов [Лопатин 2003, Григорян 1981]. Выделять ли аффиксоиды, и если выделять, то на основании каких параметров, какими свойствами должны обладать морфемы, чтобы им должен был быть приписан соответствующий статус, — проблема нерешенная.

Представляется, что, описывая эти единицы в терминах традиционных классов, все префиксоиды следует разделить на префиксы и связанные корни, а суффиксоиды — на суффиксы и связанные корни в зависимости от того, сохраняет ли та или иная единица возможность самостоятельного употребления не в составе сложного слова. Действительно, при внимательном рассмотрении класс аффиксоидов оказывается достаточно неоднородным. Возьмем, например, такие «префиксоиды», как *мега-* и *авиа-*. С одной стороны, обе эти морфемы обладают широкой сочетаемостью и достаточно четко определимым значением, но с другой, *авиа-*, в отличие от *мега-*, может выступать и как самостоятельный корень, например, в словах *авиация*, *авиатор*, да и значение его кажется гораздо более близким к обычному лексическому, нежели к значению, выражаемому словообразовательными средствами (т. е. грамматическому значению в широком понимании этого термина), что заставляет признать *авиа-* связанным корнем, а слова типа *авиастроительный* — сложными. В то же время морфема *мега-* таких свойств не обнаруживает и может быть без сомнения отнесена к классу префиксов.

Непосредственно с первой описанной проблемой связана другая, а именно — статус того или иного конкретного морфа в словообразовательной системе языка. Для создания словообразовательного словаря и разметки корпуса требуется классификация морфов, то есть нужно иметь точные списки, включающие в себя максимально большое число морфов языка, и знать, какие пометы им присваивать.

Отдельным вопросом является статус морфа не только в общей словообразовательной системе, но и внутри конкретной леммы. Действительно, в ходе исторического развития языка многие суффиксальные, а иногда и префиксальные производные от того или иного корня утрачивают свою непосредственную семантическую связь с ним и перестают ощущаться носителями языка как производные (ср. известную пару *пить — пир*). В результате процесса опрощения появляются новые словообразовательные гнезда, уже не ассоциирующиеся с теми, к которым входящие в них слова принадлежали исторически. Но где проходит граница между «еще однокоренные» и «уже не однокоренные»? Критерий «ощущаемости/неощущаемости связи носителями языка» носит достаточно субъективный характер. Разные же словари используют разные подходы: словарь Тихонова — строго синхронный, то есть префиксы и аффиксы не выделяются во всех случаях, где связь между производным и производящей основой «не ощущается» (не очевидна), словарь А. И. Кузнецовой и Т. Ф. Ефремовой зачастую обращается к диахронии, ориентация на семантику в нем минимальна. Следуя за авторами Словаря морфем, при проведении морфемного анализа мы отдаем предпочтение формальному, а не семантическому критерию определения строения слов, используя, таким образом, не только синхронный, но и исторический подход. В пользу такого решения говорит и то соображение, что «стереть» морфемную границу в спорных случаях практически всегда гораздо проще, чем провести, и в связи с этим представляется правильным предоставить более сложный вариант членения слова, который не всегда может быть получен путем интроспекции. Так, пользователь Национального корпуса русского языка, обнаруживший для слова *навзничь* разметку вида *на-взничь* и отсылку к наречию *ниц*, может без всяких дополнительных усилий счесть соответствующие два слова несвязанными и, соответственно, выделение в первом из них корня *-ничь* неоправданным; напротив, при отсутствии в Корпусе указания на такую связь пользователь, предположивший ее существо-

вание, вынужден будет обратиться к этимологическим словарям («Словарь морфем русского языка», в котором существование такой связи также признается, малоизвестен и труднодоступен).

С другой стороны, существует «обратная» проблема — проблема осложнения основ и процесс переосмысления словообразовательных связей. Так, в современном языке слово *дотошный*, восходящее к слову *точный* и исторически имеющее корень *точ*, у большого количества носителей ассоциируется в первую очередь со словом *тошнить* и, соответственно, корнем *тошн-*, а слово *столпотворение*, исторически произошедшее от слова *столп* (корень *столп-*), — со словом *толпа* (корень *толп-*).

Кроме того, в словообразовательной системе русского языка существуют так называемые поглощающие суффиксы — сращения суффиксов вида s1s2, которые требуют (для правильного предсказания акцентуации и ряда других свойств производных) разложения основ вида as1s2 на a+s1s2, а не на as1+s2, даже если слово с основой as1 существует, а слово с основой a — нет, ср. [Зализняк 1985: 60–61]. Примером поглощающего суффикса может служить сращенный суффикс *-чат-* в слове *перепончатый*: с точки зрения морфологии в этом слове выделяются два суффикса (*-к-* и *-ат-*): слова **перепона* не существует, — а с точки зрения словообразования — один (*-чат-*): в силу своих акцентных свойств суффикс *-ат-*, в отличие от суффикса *-чат-*, не может сочетаться с приставочными основами. Такие явления наводят на мысль о необходимости введения двойной разметки, которая отражала бы возникающую в подобных случаях реальную неоднозначность.

Самыми сложными для практического решения проблемами при разработке формы и параметров разметки, а также при разработке схемы ее автоматизации стали алломорфическое варьирование и омонимия аффиксов. Эти особенности словообразовательной системы русского языка не позволяют сделать процесс морфемного членения полностью автоматизированным и ставят вопрос о целесообразности их непосредственного отражения в корпусной разметке. С другой стороны, введение разметки такого уровня подробности значительно расширило бы поисковые возможности и вместе с тем — количество теоретических вопросов, которые можно было бы изучать с помощью такого инструмента, как корпус. В связи с тем, что схема автоматизации находится только в стадии разработки и совершенствования и списки морфов, составленные для упрощения работы программы-разметчика, не являются окончательными, мы решили на первом этапе отказаться от попыток разрешения проблемы омонимии аффиксов и от морфонологического компонента словообразовательной разметки и временно признать каждый алломорф отдельной единицей. Тем не менее, задача сведения алломорфов в морфемы и различения омонимичных морфем представляется одним из самых важных и перспективных направлений дальнейшей работы.

4. Порядок разметки

Словообразовательная разметка корпуса предполагает работу с полным словником НКРЯ, за исключением отдельных редких и окказиональных слов. Поскольку словник НКРЯ, который все еще находится в процессе составления, значительно превышает по объему даже словник Грамматического словаря А. А. Зализняка [Зализняк 1977/2003] (около 110 тысяч лемм), то первоначально возникла идея взять за основу большой морфемный или словообразовательный словарь, который затем можно было бы дополнить в части новых слов.

На первый взгляд, хорошим претендентом на эту роль кажется словарь А. Н. Тихонова [Тихонов 1985], который обладает чрезвычайно большим объемом словника (154 000 единиц). Однако этот словарь не дает, на наш взгляд, удовлетворительного морфемного анализа в большинстве спорных и сложных случаев (ср. *сказ-к-а, поезд* и др.), и это делает его использование в качестве основы для разметки малоприемлемым.

«Словарь морфем русского языка» [Кузнецова, Ефремова 1986], включает всего 52 000 лемм — то есть, если использовать только этот словарь для разметки текстов, то больше половины слов корпуса не получат в таком случае словообразовательных помет. Вместе с тем, сам словарь устраивает нас в том отношении, что в решении основного круга теоретических вопросов мы следуем за его авторами. В результате было принято решение создать собственный морфемно-словообразовательный словарь корпуса, взяв при этом за основу Словарь морфем: использовать представленные в нем списки аффиксов и алломорфов, а также привлечь данные словаря для разрешения сложных случаев морфемного членения, в частности, при выделении суффиксов. Вместе с тем, был также составлен список поправок, касающихся некоторых конкретных решений Словаря морфем, которые показались нам неприемлемыми. Например, Словарь морфем выделяет в слове *судья* аффикс *ья* (по нашему мнению, в орфографической записи этот суффикс должен иметь вид ь), в словах типа *подготавливать* Словарь морфем выделяет *л* как отдельный суффикс, идущий после корня *готав* (мы же считаем его входящим в состав корня *готавл*, чередующегося с *готов*, ср. пару *подготовить — подготавливать*) и т. д.

Итак, поскольку Словарь морфем покрывает лишь небольшую часть словника НКРЯ, речь фактически идет о самостоятельной разметке нового словаря корпуса силами сотрудников проекта (есте-

ственно, с привлечением данных словарей на тех участках, где это возможно). Ручная обработка словника подобного объема с предполагаемой нами степенью подробности представляется задачей трудновыполнимой. В связи с этим единственным возможным решением представляется разработка схемы автоматизации морфемного анализа, которая позволила бы проделать большую часть работы по отделению аффиксов в автоматическом режиме. К сожалению, создать точный автоматический разметчик практически невозможно, и полностью избежать вмешательства исследователя в процесс морфемного анализа не удастся, но все же «ручную» часть можно свести к минимуму, заключающемуся только в проверке результатов работы программы, если осуществить хотя бы первичное деление и приписывание морфов в автоматическом режиме.

По ряду причин наименее проблемной зоной для автоматической разметки оказываются префиксы. Во-первых, префиксальные морфы достаточно легко отделяются даже в орфографической записи, проблема проведения морфемной границы (актуальная, например, для глагола *обрыднуть*) возникает в ничтожно малом количестве случаев. Во-вторых, префиксы обладают в общем случае достаточно широкой сочетаемостью, что значительно упрощает разработку алгоритма по их отделению. Большое преимущество задачи отделения префиксов состоит в том, что для них возможно разработать алгоритм автоматической разметки без опоры на другие аффиксы, чего нельзя осуществить ни для корней, ни для суффиксов. К тому же, алломорфическое варьирование не так распространено в зоне префиксов, как в зоне суффиксов, что значительно упрощает решение проблемы сведения алломорфов в морфемы. Одним из главных факторов, облегчающих выделение префиксов, является то, что префиксы образуют «кластеры» гораздо реже, чем суффиксы, и среди префиксальных кластеров фактически не встречается неделимых (по крайней мере, неделимых в орфографической записи).

Разметка суффиксальной части лемм очевидно должна вызвать большое количество проблем в связи с широко распространенным в этой части алломорфическим варьированием и со сращениями суффиксов. Разметка суффиксов будет проводиться с опорой на Словарь морфем А. И. Кузнецовой и Т. Ф. Ефремовой [Кузнецова, Ефремова 1986], а также на работу [Иткин 2007], содержащую наиболее подробное на сегодняшний день описание алломорфического варьирования всех аффиксальных морфем русского языка, кроме заимствованных. Для разметки слов, не вошедших в Словарь морфем, будут построены алгоритмы-эвристики, учитывающие опыт разборов в Словаре. Отдельную задачу составит составление инвентаря аффиксов в заимствованных словах и разметка заимствованных слов — эта задача Словарем морфем не решается.

Без сомнения, самой сложной проблемой при автоматическом морфемном анализе представляется отождествление корней, так как полных списков словообразовательных гнезд, удовлетворяющих нашим целям, не существует. Отделение префиксов и суффиксов должно значительно облегчить задачу по выделению и — частично — по отождествлению корней. Таким образом, правильным порядком разметки представляется следующий: отделение префиксов, затем суффиксов, отождествление корней.

5. Мы делили *а-пельсин*, или как отделить префиксы

Первой задачей на пути морфемного анализа словника НКРЯ стала задача автоматического отделения префиксов. Эта работа включала в себя несколько теоретических и практических этапов, в том числе: составление списков префиксальных морфов, которые послужили бы основой работы программы-разметчика, разработка схемы автоматизации, написание программы-разметчика в соответствии с разработанной схемой, ручная проверка результатов работы программы, попытка оптимизации работы программы по результатам ручной обработки размеченного словника.

5.1. Вспомогательные списки

Для того чтобы автоматически отделить префиксы и выделить первые корни сложных слов, необходимо было составить как можно более полные списки префиксальных морфов и связанных частей сложных слов, на которые могла бы опираться программа-разметчик. Несмотря на существующие в разных источниках списки подобного рода, процесс составления полных списков оказался достаточно трудоемким: список префиксальных морфов, данный в Русской Грамматике [РГ 1980], оказался неполным, а список «повторяющихся (в том числе связанных) частей сложных слов» был составлен достаточно непоследовательно. Принцип, по которому авторы выбирали вошедшие в список «повторяющиеся части» из всей массы встречающихся в начале сложных слов, совершенно неясен. Неполнота (с нашей точки зрения) списка префиксальных морфов объясняется общим подходом авторов к морфемному анализу: редкие префиксы, чьи немногочисленные производные претерпели ряд семантических изменений и утратили прозрачную связь с мотивирующей основой (а иногда — и саму мотивирующую основу), авторами [РГ 1980] не выделялись и в список не вошли. Так, например, в Русской Грамматике отсутствует префикс *ку* (*кумекать, скукожиться*), выделяемый, однако, Словарем морфем Кузнецовой

и Ефремовой [Кузнецова, Ефремова 1986]. Изъяном же списка префиксов, достаточно последовательно представленного в самом Словаре морфем, является принципиальное отсутствие в нем заимствованных морфов, число которых в русском языке достаточно велико и морфологический статус которых представляет зачастую отдельную проблему. Таким образом, списки, ставшие основой работы нашей программы-разметчика, были составлены на основе списков Словаря морфем [Кузнецова, Ефремова 1986] и Русской Грамматики [РГ 1980]. Элементы, вошедшие в список «повторяющихся (в том числе связанных элементов сложных слов)» Русской Грамматики, которые все вместе могли бы претендовать на статус так называемых аффиксоидов, были расклассифицированы на основании их деривационных свойств в две группы: префиксальные морфы и «корни, связанные справа». Таким образом, из используемой нами в разметке системы терминов был исключен термин «префиксоид».

5.2. Схема автоматизации разметки

Несмотря на то, что существуют программы, производящие автоматический морфемный анализ слов, программы, отделяющей только префиксы и анализирующей сложные слова, по нашим сведениям не существует. В то же время, разделение этапов морфемного анализа слова представляется достаточно правильным подходом: проверка результатов разбора после отделения морфем одного типа позволяет избежать накопления ошибок, неизбежного при одновременном полном анализе слов: поскольку морфы в слове непосредственно контактируют, проведение одной неверной морфемной границы ведет к приписыванию слову минимум сразу двух неверных помет, то есть ошибка на стадии выделения префикса автоматически влечет за собой неправильное выделение корня.

В связи с этим перед нами встала задача разработки «с нуля» схемы автоматизации отделения префиксальных морфов. Основным свойством словообразовательной системы русского языка, которое позволило достаточно успешно осуществить поставленную задачу, является то, что в большинстве случаев префиксальные производные имеют в языке соответствующую беспрефиксную пару, послужившую для них производящей основой.

5.3. Принцип работы программы

Наша программа реализована на базе технологии .NET Framework, в среде Microsoft Visual Studio.

На вход программе подается список лемм (в данном случае — словник Грамматического словаря). Опираясь на списки префиксов и связанных частей сложных слов, программа проверяет наличие соответствующих начальных частей в леммах словника по порядку убывания длины морфа: сначала осуществляется поиск соответствий более длинным приставкам, затем более коротким, с целью уменьшения количества ошибок (например, чтобы в словах с префиксом *между* не был выделен префикс *меж*). Затем программа проверяет наличие в словаре леммы, соответствующей неприкрытой части леммы с выделенным префиксом. Если таковая существует, лемме приписывается свойство «имеет префикс» и указываются префикс и неприкрытая часть основы. Если находятся две или более цепочки букв разной длины (как в случае *меж* и *между*), соответствующие разным префиксам списка (или префиксу и связанному корню), программа приписывает оба возможных разбора (для разрешения подобных спорных случаев необходима ручная проверка результатов работы программы). Процедура повторяется для всех префиксов списка. После завершение первого круга проверки, программа тем же образом проверяет отделенные неприкрытые части лемм, выделяя таким образом не только первые, но и вторые, и последующие приставки. Вся процедура занимает около 4 минут. Зависит это время от объема анализируемого словника. На выход подается таблица с разборами для каждого слова.

Естественно, при использовании описанной выше схемы неизбежным является получение достаточно большого количества неверных разборов, в которых морфемная граница проведена там, где ее на самом деле не существует (ср., например, получившиеся в результате разметки разборы *бес-еда* и *на-гайка*). Достаточное количество подобных разборов также делает необходимой ручную проверку результатов работы программы.

На финальном этапе работы автоматический разметчик проверяет список лемм, не получивших разбора, на наличие связанных корней: ищет одинаково оканчивающиеся леммы, и если начальные части таких лемм являются приставками, входящими в составленный список префиксов, то им приписываются свойства «имеет префикс» и «имеет связанный корень».

Разработанная программа осуществляет также предварительную обработку сложных слов, опираясь на список связанных частей сложных слов (таким образом выделяются сложные слова со связанными начальными корнями) и на следующий принцип: если в списке лемм, не имеющих префиксов, есть леммы, оканчивающиеся на части, идентичные другим леммам из данного списка, а начальные части не соответствуют префиксам и связанным корням, то таким леммам приписывается первичная помета «сложное слово».

Как уже было сказано выше, в связи с предусмотренной возможностью приписывания нескольких разборов и с тем, что принцип работы программы

все-таки несовершенен и не дает стопроцентной точности в разборе, результаты работы программы нуждаются в постредактировании, осуществляемом одним или несколькими людьми.

5.4. Результаты работы программы

Разработанная нами по вышеописанному принципу программа работает с точностью, приблизительно равной 90%, что представляет собой достаточно высокую степень точности, учитывая количество нерегулярных случаев словообразования в русском языке. В связи с предусмотренной возможностью нескольких вариантов разборов одной леммы в результате ее работы для 110 000 лексем, входящих в словник НКРЯ, было получено примерно 125 000 возможных разборов. Таким образом, ручная проверка результатов работы программы, как и предполагалось заранее, оказалась неизбежной.

Для оптимизации и ускорения процесса ручной проверки результатов была создана специальная компьютерная программа — рабочее место постредактора. Общий список лемм был разбит на равные части (приблизительно по 20 000 лемм каждая), каждая из которых проверялась отдельно разными участниками проекта. Спорные случаи разбора, а также статус отдельных морфов обсуждались совместно. После первичной проверки отдельные отредактированные части были вновь собраны в единый массив и подвергнуты вторичной проверке на предмет единообразия принятых по спорным случаям решений.

В процессе ручной обработки результатов работы программы (постредактирования) было выявлено несколько проблем. Во-первых, составленные нами списки, служившие основой работы программы, оказались неполными в части «связанных корней сложных слов», что неудивительно, учитывая их количество, а также тот факт, что предварительных полных списков «первых частей сложных слов» у нас в распоряжении не было, и приходилось работать, как описано выше, с достаточно непоследовательно составленным и кратким списком, представленным в [РГ 1980]. С другой стороны, списки префиксальных морфов оказались «слишком полными»: входящие в них редкие префиксы были неверно отделены программой в очень большом количестве случаев (это касается, в первую очередь, префикса *к-*, выделяемого только в нескольких случаях — в наречиях *кверху* и *книзу* и нек. др.; ср. неправильные случаи членения *к-лад, к-рот* и т. п.). Это заставило нас включить в программу не только списки префиксов, но и списки всех производных для каждого из редких префиксов (а именно, префиксов с не более чем десятью производными), чтобы исключить лишние случаи отделения подобных редких морфем. Кроме того, в результате постредактирования на основании получившихся результатов вручную были пополнены списки «связанных корней сложных слов», что также позволило увеличить процент точности, с которым работает программа.

6. Заключение

Проект рассчитан на три года. Несмотря на то, что значительная часть работы уже проделана (разработаны общая концепция и порядок разметки, а также разработана и написана программа по автоматическому выделению префиксов), в ближайшие два года нам предстоит решить еще довольно большое количество сложных теоретических и практических задач. К сожалению, до сих пор не разработана схема автоматизации разметки суффиксов и корней. К тому же одним из по-прежнему актуальных направлений работы является усовершенствование программы по отделению префиксов с тем, чтобы повысить точность ее работы и свести постредактирование («ручную» часть) при разметке добавлений к словнику Грамматического словаря Зализняка (словника НКРЯ) к минимуму.

Одной из самых главных задач как с точки зрения практики (то есть разметки), так и с точки зрения теории станет составление списков морфем с их алломорфами (то есть сведение префиксальных и аффиксальных алломорфов в морфемы) и разрешение омонимии аффиксов (то есть составление списков морфем с различением омонимичных аффиксов). Эта задача представляется очень сложной с теоретической точки зрения и в связи с этим достаточно трудоемкой. Кроме того, благодаря существованию различных подходов к решению вышеописанных теоретических вопросов многие случаи неизбежно должны вызвать большое количество споров. Но, несмотря на все это, подобные списки станут большим шагом на пути к подробному описанию системы словообразовательных единиц, заполнив важные лакуны в существующих ныне описаниях, таких, как «Словарь морфем русского языка» А. И. Кузнецовой и Т. Ф. Ефремовой [Кузнецова, Ефремова 1986].

Литература

1. *Rasch, Barbara J. L.* A syntactic analysis of word-formation in Russian with particular emphasis on deverbal nouns. PhD dissertation. University of Washington, 1977.
2. *Townsend, Charles E.* Russian word-formation. Slavica publishers, 1968.
3. *Wolkonsky C., Poltoratzky M.* Handbook of Russian Roots. London — New York: Columbia University Press, 1969
4. *Виноградова В. Н.* Стилистический аспект русского словообразования. М.: Наука, 1984.
5. *Григорян Э. А.* Суффиксоиды в системе современного русского языка (на материале сложений со вторым глагольным компонентом). Диссертация на соискание ученой степени кандидата филологических наук. М., 1981.
6. *Ефремова Т. Ф.* Новый толково-словообразовательный словарь русского языка. М., 2000.
7. *Зализняк А. А.* Русское именное словоизменение. М., 1967.
8. *Зализняк А. А.* Грамматический словарь русского языка. М.: Русский язык, 1977. 4-е изд. М.: Русские словари, 2003.
9. *Зализняк А. А.* От праславянской акцентуации к русской. М.: Наука, 1985.
10. *Земская Е. А.* Современный русский язык. Словообразование. М., 1973.
11. *Земская Е. А. Словообразование как деятельность. М., 1992.*
12. *Изменения* в словообразовании и формах существительного и прилагательного в русском литературном языке XIX века. Очерки по исторической грамматике русского литературного языка XIX века, под ред. В. В. Виноградова и Н. Ю. Шведовой. М., 1964.
13. *Иткин И. Б.* Русская морфонология. М., Гнозис, 2007
14. *Кубрякова Е. С.* Что такое словообразование. М., 1965.
15. *Кузнецова А. И., Ефремова Т. Ф.* Словарь морфем русского языка. М., 1976.
16. *Лопатин В. В.* Аффиксоид // Русский язык. Энциклопедия / Под ред. Ю. Н. Караулова. М., 2003.
17. *Пазельская А. Г.* Модели деривации отглагольных существительных: взгляд из корпуса // Корпусные исследования по русской грамматике. М.: Пробел-2000, 2009.
18. *Плунгян В. А.* Общая морфология: Введение в проблематику. М.: Эдиториал УРСС, 2000
19. *Потиха З. А.* Школьный словообразовательный словарь (под ред. С. Г. Бархударова). М.: 1961.
20. *Развитие* современного русского языка. 1972. Словообразование. Членимость слова. М., 1975.
21. *РГ 1980:* Русская грамматика. М.: Наука, 1980.
22. *Тихонов А. Н. Словообразовательный словарь русского языка. В двух томах. М.: Русский язык, 1985.*
23. *Тихонов А. Н.* Школьный словообразовательный словарь русского языка. М., 1978.
24. *Улуханов И. С.* Единицы словообразовательной системы русского языка и их лексическая реализация. М.,1996.
25. *Улуханов И. С.* Словообразовательная семантика в русском языке и принципы ее описания. М., 1977.
26. *Шанский Н. М.* Очерки по русскому словообразованию. М., 1968.
27. *Шанский Н. М., Тихонов А. Н.* (ред.). Современный русский язык: В 3-х ч. Словообразование. Морфология. М., 1981. Ч. 2.

Лексико-грамматические базы данных и сравнительное изучение русских диалектных акцентных систем[1]

Lexico-grammatical databases and comparative study of Russian dialectal accentuation systems

Тер-Аванесова А. В. (teravan@mail.ru)
Институт русского языка им. В. В. Виноградова РАН

Крылов С. А. (krylov-58@mail.ru)
Институт востоковедения РАН, Институт системного анализа РАН

С помощью СУБД StarLing обогащена ранее построенная база данных по русским говорам с различением двух фонем «типа *о*». Включение в нее данных нескольких говоров позволило провести сравнение их акцентных систем. Проведена акцентологическая разметка базы данных.

1. Лексико-грамматические базы данных по двум русским говорам — среднерусскому говору с. Пустоша Шатурского р-на Московской обл. (далее Пуст.) и севернорусскому слободскому говору дд. Арзубиха, Захариха и Злобиха Харовского р-на Вологодской обл. (далее Сл.) — представляют собой полные описания этих говоров в рамках имеющихся корпусов текстов. Поскольку корпусы текстов содержат не только расшифровки аудиозаписей «спонтанной» («связной») речи носителей говоров, но и ответы на обширные вопросники, ориентированные на сбор морфологических, акцентологических, фонетических данных, то даже небольшие по объему корпусы достаточно полно охватывают исконную лексику говора и демонстрируют все существенные словоизменительные и акцентуационные классы слов.

Помимо названных, были созданы базы данных южнорусских говоров с. Новоселки Рыбновского р-на Рязанской обл. (Нов.), задонского (дд. Гнилуша и Воскресеновка Задонского р-на Липецкой обл.) и калужского (дд. Пеневичи, Бояновичи, Кудрявец Хвастовичского р-на Калужской обл.) (Пен.).

В результате появилась возможность сравнения говоров с помощью СУБД StarLing.

Выбор говоров обусловлен тем, что все они имеют так называемый «семифонемный» ударный вокализм: в них различаются две фонемы «типа *о*» и две фонемы «типа *е*». Фонетические реализации фонем «типа *о*» и «типа *е*» в говорах в общем виде можно описать так: фонемы /ω/ («*о* закрытое») и /ие/ («*е* закрытое») представлены монофтонгами верхне-среднего, реже — среднего подъема и дифтонгоидами типа *уо*, *ие*, фонемы /о/ («*о* открытое») и /е/ («*е* открытое») представлены монофтонгами средне-нижнего подъема и дифтонгоидами типа *оу*, *еи*. Этот тип вокализма в русских говорах довольно редок; однако благодаря ему говоры приобретают особую историческую и типологическую значимость.

Различие фонем «типа *е*» и «типа *о*», как известно, обусловлено исторически: /ие/ восходит к праслав. **ě*, /е/ — к **е* и **ь*. Распределение двух *о* в русском описывается правилом: фонема /ω/ выступает на месте **о* под автономным ударением, фонема

[1] Авторы благодарят В. И. Беликова за высказанные им замечания к статье и А. С. Касьяна за помощь в подготовке текста. Работа выполнена при поддержке Программы АИФН РАН «Генезис и взаимодействие социальных, культурных и языковых общностей», проект «Восточнославянский диалектный корпус: праславянское наследие и лингвогеография» и проект «Генезис балто-славянской языковой общности: акцентологический аспект».

/о/ — на месте **о* под автоматическим ударением[2], на месте **ѣ*, **е*, **ь* [Зализняк 1985: 173–179]. Правило распределения фонем /ω/ и /о/ в русском было установлено А. А. Шахматовым [Шахматов 1914] и Л. Л. Васильевым [Васильев 1929] и сформулировано ими в терминах слоговых интонаций.

Правило Васильева — Шахматова устанавливает связь современных рефлексов **о* с характером праславянского (или древнерусского) ударения словоформ, в фонемный состав которых эти гласные входят. В сфере действия правила тем самым находятся словоформы исконных слов, ударение которых сохраняет свое старое место (и заимствований, вошедших в русский язык до того, как автономное и автоматическое ударение совпали). В качестве такого класса слов в русском языке обычно рассматривались непроизводные существительные муж. рода. Тот или иной тембр корневого *о* словоформ, ударение которых в истории языка изменило свое место, более поздних заимствований и новообразований, обусловлен иными закономерностями, нежели правило Васильева — Шахматова. Между тем зависимость рефлексов **о* от характера праславянского ударения никогда не рассматривалась на достаточно большом материале слов какого-либо класса. Многочисленные примеры выравнивания огласовки корней заставляют опасаться, что и в обычно считавшихся архаичными категориях случаев также могли происходить процессы выравнивания.

С помощью СУБД StarLing проведено сравнение акцентуации непроизводных слов во всем имеющемся корпусе материалов с целью выяснить степень близости систем разных говоров и установить, какая связь существует между огласовкой корней, содержащих **о*, с синхронной акцентуацией слов (синхронными диалектными акцентными типами) и древнерусской (праславянской) акцентуацией тех же слов (их праслав. акцентными парадигмами).

В лексико-грамматической базе данных информация о словоформе включает:

(76) словоформу в фонемной транскрипции;
(77) грамматическую помету;
(78) акцентный тип (далее а. т.);
(79) соответствие праславянской акцентной парадигме (далее а. п.).

Разметка базы данных позволяет сравнивать друг с другом:

(1) акцентные контуры, представленные в разных говорах у данной словоформы;
(2) акцентные кривые, представленные в разных говорах у данной лексемы;
(3) списки слов, входящих в разных говорах в данный акцентный тип;
(4) списки слов различных синхронных акцентных типов, относящихся к данной праславянской акцентной парадигме.

Базы данных позволяют тем самым выполнить акцентную реконструкцию на имеющемся диалектном материале и соотнести ее с праславянской акцентной реконструкцией. Базы данных в формате StarLing позволяют получить различными способами упорядоченные списки слов (словоформ) с корневым ударным *о* того или иного тембра, снабженных пометами, относящимися к разным уровням акцентологического описания.

Ниже приводятся основные результаты исследования на материале существительных муж. рода.

2. Наборы «стандартных» акцентных типов (акцентных кривых) существительных муж. рода в говорах одинаковы; они показаны в таблице 1. Кроме того, в Сл. имеется особый а. т. C1 у четырех основ на *-ень* (*гре́бень*, *ре́мень*, *сли́зень*, *че́рень;* противопоставлен а. т. C, к которому относятся *блу́день*, *во́тень*, *ка́мень*, *ко́рень*, *ко́чень*, *кре́мень*, *лёжень*, *па́рень*, *пе́рстень*, *сте́ржень*, *шкво́рень*); от «стандартного» а. т. C он отличается начальным ударением формы gen.pl. Как особый а. т. C2 можно рассматривать акцентные кривые, отличающиеся от а. т. C нафлексионным ударением счетной формы (например, gen. *гу́ся*, num. *гуся́*, in. *гу́сём*, и т. д.). К а. т. C2 в говорах относится по два-три слова, и лишь в Сл. их семь (*гусь*, *груздь*, *клок*, *раз*, *рог*, *шаг*, *час*). Некоторые слова имеют индивидуальные акцентные кривые, наиболее важные из них рассмотрены ниже.

3. Распределение лексем муж. рода по а. т. в говорах различается деталями, в целом совпадая с тем, что представлено в литературном языке. У отдельных слов отмечается колебание а. т. (*творог*, *под* в Пуст., *потолок* в Сл., *хворост* в Нов. и др.).

Соотношение рефлексов **о* в корнях существительных муж. рода с а. т. (A, B, C, D) последних, наблюдаемыми в говорах, показано в таблице 1[3].

---

[2] Фонетическое различие *автономного* (фонологически значимого) ударения и *автоматического* (фонологически тождественного безударности) предположительно существовало в древнерусском языке; впоследствии оба ударения совпали в едином современном ударении, см. [Зализняк 1985: 119–121]. Автономное ударение было характерно для *ортотонических* словоформ, автоматическое ударение — усиление начального слога словоформы или проклитико-энклитической группы в определенных фразовых условиях — было характерно для *форм-энклиноменов* [Дыбо 2000: 25–30]. Считается, что автономное ударение реализовалось в *«восходящей»*, а автоматическое — в *«нисходящей»* интонации ударного слога; уточнения этого положения см. в [Дыбо 2000: 27–28].

[3] Слова, имеющие в говорах одинаковую акцентуацию и одинаковый фонемный состав, приводятся без географических помет. Они даны в условной записи, близкой к фонемной. Варианты /ω/ передаются буквой ω (*стωл*), варианты /о/ — буквой *о*.

Таблица 1

	A	B		C	D	
nom. sg.	*морóз* *колóдезь*	*кωт* *кωнь*	*плод* *гвоздь*	*год* *гость*	*бор* *рой*	*бωй*
gen. sg.	*морóза* *колóдца*	*котá* *коня́*	*плодá* *гвоздя́*	*гóда* *гóстя*	*бóра* *рóйа*	*бωйа*
loc. sg.	*морóзе* *колóдце*	*котиé* *конúе*	*плодиé* *гвоздиé*	*гóде/годý* *гóсте*	*борý* *ройý*	*бойý*
nom. pl.	*морóзы* *колóдцы*	*коты́* *конú/кóни*	*плоды́* *гвóзди*	*гóды/годá* *гóсти*	*боры́/борá* *роú*	*боú*
gen. pl.	*морóзов* *колóдцев/-ей*	*котóв* *конéй*	*плодóв* *гвоздéй*	*годóв* *гостéй*	*борóв* *ройóв*	*бойóв*

Примечания к таблице. А.т. C и D противопоставлены только в формах loc.sg. и n.-acc.pl. с окончанием *-ы/-и*. В n.-acc.pl. это противопоставление у многих слов нейтрализовано в результате распространения ударного окончания *-á*, которое во всех говорах, а особенно в Нов. и Пуст., встречается у значительно большего числа слов, чем в лит. языке.

К а. т. C отнесены слова с начальным ударением в формах ед. числа и в форме n.-acc. pl., имеющей окончание *-ы/-и*, и ударением на окончании в прочих формах мн. числа. Многие слова а. т. C с основами на твердый согласный и *j* имеют ударное окончание loc.sg. *-ý* (наряду с безударным окончанием *-е* или как единственно возможное) и ударное окончание n.-acc. pl. *-á* (наряду с безударным окончанием *-ы/-и* или как единственно возможное). Все слова с ударным окончанием n.-acc.pl. *-ы́/-и́*, а также слова, у которых наряду с *-ы́/-и́* отмечено окончание *-á*, отнесены к а. т. D. У слов а. т. D в loc.sg. встречается только окончание *-ý*.

Таблица показывает, что непроизводные существительные муж. рода на месте **o* имеют в корнях фонему /ω/, если они относятся к а. т. A, и фонему /o/, если они относятся к а. т. C. Слова а. т. B и D, делятся на две группы: с /ω/ и /o/ в корне. Огласовка корня, таким образом, не всегда напрямую соотнесена с синхронным акцентным типом слова.

4. Соотношение огласовки корней существительных муж. рода с праславянской акцентуацией последних выглядит следующим образом.

4.1. Праслав. а.п. *a* сохраняется практически без изменений, отражаясь в **а. т. A**. Словоформы а. т. A ведут себя как ортотонические: их основы содержат под ударением /ω/ < **o* и у них отсутствует перенос ударения на приставки.

К а. т. A относятся слова, у которых в праславянском реконструируется «старый акут»: *горóх*, Пуст. *пъддорóжник*, Сл. *зорóд* ‘сушило для сена’, Пуст.Нов. *колóдезь* (gen. *колóдца*), *морóз*, *порóг*, *порóм* (но Пуст. *порóм* — неосвоенное слово?).

Свойствами ортотонических словоформ обладают формы приставочных и приставочно-суффиксальных девербативов с неподвижным ударением на корне: *зверобóй*, *забóр*, *убóр*, *завóд*, *полёвóд*, *свóд*, *(у)повóйник*, *навóз*, *завóр* (Сл. ‘раздвижные ворота’, Пуст. ‘калитка’), *поворóт*, *розговóр*, *сгóн* ‘выгон скота в стадо’, *самогóн*, *прогóн* ‘переулок’, Сл. *огорóд* ‘изгородь’, Сл. *повóст* ‘погост’, *удóй*, *подóйник*, *пустозвóн*, *(над-, под-)зóр*, *(у)покóйник*, *укóл*, Сл. *тынокów* (вид изгороди), Пуст. *покóн* ‘канун праздника’, *(по-, сено-)кóс*, *укрóп*, *улóв*, *налóг*, *запóр*, *(на-, у-)рóд*, Сл. *нарóст* ‘поколение’, *подрóсток*, Сл. *суслóн* ‘укладка ржаных снопов’, *засóл*, Пуст. *отстóй* ‘осадок’, Сл. *росстрóй* ‘огорчение’, *затвóр*, *потóк*, Сл. *потóшник* ‘водосточный желоб’, *востóк*, *подтóпок*, *(в-, об-, по-, при-, про-)хóд*, Пуст. *потолóк*, gen. *потолóка* (*/потолóк*, gen. *потowká*, а. т. B). Эта словообразовательная и акцентуационная модель появляется уже в праславянском, но, как видно из списка, в говорах в нее включены и более поздние образования. То же относится к большинству отыменных производных а. т. A: *подборóдок*, *ворóбушек*, *втóрник*, *колоколóльчик*, *робóтник*, *рóдственик*, *срóдник*, *охóтник*, *сапóжник*. Исключением являются относящиеся к а. т. A диминутивы с суффиксами *-ик*, *-ышок*, огласовка корня которых повторяет огласовку корня производящих слов: *кóлышок*, *кóтик*, *нóжык*, *стóлик*, *хвóстик* (ср. *кωл*, *кωт*, *нωж*, *стωл*, *хвωст*), *гвóздик*, *гóдик*, *грóбик*, *дóмик*, Сл. *кóчик*, *лóмик*, *нóсик*, *снóпик*, *пóлик* (ср. *гвоздь*, *год*, *гроб*, *дом*, *коч*, *лом*, *нос*, *сноп*, *пол*).

4.2. Праслав. а. п. *b* отражается в **а. т. B** также практически без исключений; все словоформы а. п. *b* ортотонические, что согласуется с развитием корневого **o*>/ω/ под ударением: *батóг*, *двор*, *жывóт*, *кωл*, *кωнь*, *кωт*, *нωж*, *пирóг*, *пωп*, *скωт*, *стωл*, *ствωл*, *топóр*, *хвωст*, Пуст.: *бωб*, *клωп*, *корóль*, *пωд* (*/под*, gen. *пóда*), *снωп*, *творóг*, *хвωщ*, Нов.: *клωп*, *мазóль*, *пωд*, *снωп*, *хварóст*, (*/хвóраст*, gen. *хвóръсту*), Сл. *пωст*.

Исключением из правила Васильева — Шахматова является развитие **ъ* >/ω/ в Пуст.Сл. *крωт* (но Лек.Нов. *крот*).

4.3. Рефлексы праслав. а.п. *c* и *d* показаны в таблице 2. Они зависят от старого типа основы существительного, а у **u*-основ — от праслав. количества корневого гласного. Характер рефлексов а. п. *c* и *d* у существительных муж. рода указывает на принадлежность говоров к «восточному» типу позднепраславянских диалектов [ОСА 1990: 155–159].

Таблица 2

	*o-основы		*u-основы		*i-основы
Корневой праслав. гласный	долгий	краткий	долгий	краткий	долг., кратк.
а. п. *c*	C	C	D	C, особ. рефлексы	C
а. п. *d*	C, D, B	C, D, B, особ. рефлексы	D, B, особ. рефлексы	D, B, особ. рефлексы	C, D, B

Словоформы существительных м. рода а. п. *c* и *d* с ударением на начальном слоге должны рассматриваться как рефлексы форм-энклиноменов: у них отмечается перенос ударения на предлоги и под ударением на месте **o* выступает фонема /o/.

4.3.1. Ниже приводится список слов **а. т. C** с корневым **o*, в который также включены (на основании имеющихся у них признаков а. т. C) singularia tantum, слова с осложненной основой мн. числа и слова, у которых формы мн. числа не были зафиксированы при сборе материала.

бог, *бок*, *бор*, *бóров* ‘часть дымохода’, ‘поросёнок’, ‘толстяк’, *брод*, *воз*, *вóлок* ‘отрезок пути’, *вóлос*, *вор*, *вóрон*, *вóрот*, *вóрох*, *воск*, Сл. *гвоздь*, *гóвор*, *год*, *гóлод*, *гóлос*, *гóлубь*, Пуст. *гон* ‘брачный период у лосей’, *гóрод*, *гость*, *гроб*, *гром*, Сл. *дой*, *дом*, Сл. *дор* ‘росчисть в лесу’, *звон*, *зной* (Сл. Пуст. ‘жара’, Нов. ‘*хóлъднъ ил’ жáркъ*’), *зоб*, *зов* (Сл. Пуст. ‘брачный период у некоторых животных’), Сл. Нов. *клок*, Сл. Пуст. *кóготь*, *кóлокол*, *кóлоб* (Пуст. ‘лепешка’, Сл. ‘жмых из льняного семени’), *кóлос*, *ком*, *кон*, *кóрень*, *кóроб*, Сл. *кóчень*, Сл. *лóкоть*, *лов*, *лом* ‘орудие’, ‘сломанные вещи’, ‘ломота’, Пуст. *лось*, *мозг*, *мóлот*, *мор*, Пуст. *мóст* ‘мост; сени’, Пуст. *нóготь*, *нóкоть* ‘ноготь, коготь’, *нóров*, Сл. Нов. *нос*, *óкунь*, Сл. *óстров* ‘нехоженая часть леса’, Нов. *под*, *пол* ‘пол’ (Нов. устаревш. ‘земляной пол’), ‘половина’, Пуст. *пóломень*, *пот*, Сл. Нов. *рог*, *род*, *рост*, Нов. *плод*, *пóлоз*, *пóрох*, *сóболь*, *сок*, *сóкол*, *сóлод*, *сор*, *стог*, Нов. *стон*, *стóрож*, *твóрог*, *ток* ‘гумно’, ‘брачный период у птиц и змей’; ‘электрический ток’, *тóполь*, *хвóрост*, *хóбот* ‘часть колодца-журавля’, ‘хобот слона’, *ход*, *хóлод*. Производные а. т. C: *дóговор*, *óбод*, *óкорок*, *пóгреб*, *пóдпол*, *пóйезд*, *пóяс*, *прóтивень*, Сл. Нов. *óсек* ‘изгородь из свежесрубленных деревьев’, Пуст. Нов. *пóдмъст* ‘подпол’, Сл.: *óпорос*, *óтвод* ‘ворота’, *óцеп* ‘жердь, на которой подвешена колыбель; часть колодца-журавля’.

**u*-основа а. п. *с дом* в говоре Пуст. имеет нафлексионное ударение в форме gen.sg. в конструкциях с предлогами *из*, *от: из домá*, при *ниéд дóма*, *до дóма*; **u*-основа а. п.*c слиед* — нафлексионное ударение в gen.sg.: *слéдá*, при num. *слиéдъ*, dat. *по слиéду*, in. *слиéдъм*, nom. pl. *слéды́ ~ слéдá*, gen. *слéдώф*, т. е. оба слова обнаруживают следы акцентной кривой а. п.*c* **u*-основ [Дыбо 2000: 23].

4.3.2. Долготные и краткостные односложные **o*-основы а. п. *d*, долготные **u*-основы а. п. *d* и *c* образуют **а. т. D** с начальным ударением форм ед. числа и нафлексионным (при n.-acc.pl. *-ы́/-и́*) — форм мн. числа. Начальноударные формы слов а. т. D являются рефлексами форм-энклиноменов (/o/<**o* в корне под ударением, факультативная оттяжка ударения на предлоги):

слова с корневым **o* > /o/, Пуст.: *гроб*, *ход*, Сл.: *бор*, *брод*, *гроб*, *клок*, *лом*, *мост* ‘настил из досок: пол, мост’, *рой*, Нов.: *пол*.

4.3.3. Прямым отражением а. п. *d* является **а. т. B** нескольких **u-*, **o-*, **i-* и консонантных основ с рефлексами форм-энклиноменов в n.-acc.sg., а у **i-* и консонантных основ — также в n.-acc.pl., при нафлексионном ударении прочих форм парадигмы.

Слова с корневым **o*>/o/: *вол*, *дрозд*, *сом*, *хорь*, Сл. *гроздь*, *клоп*, *коч* ‘кочка’, *плод* (/D), *под* (/D), *сноп* (/D); Пуст. *плод*, *гвоздь*, *лóкоть*, *(под)*; Нов. *гвоздь*, *нóготь*, *лóкоть*, *лось* (/C), *мост*, *пост*.

У нескольких слов а. п. *d* представлено варьирование а. т. C/B. Пуст.: **o — рок*, num. *двá рогá*, gen. *рогá*, in. *рóгъм*, n. pl. *рогá; нос*, gen. *нóса*, ńm. *нóса*, in. *пóд нъсъм/под нóсъм /пъд носóм*, n.pl. *носá; *ъ — горп*, gen. *горбá*, instr. *гóрбъм*, loc. *нъ горбý*, nom. pl. *горбá*. Нов. *лос’*, num. *два лас’á*, gen. *лóс’а ~ лас’á*, in. *за лóс’ам*, n.pl. *лóс’и;* с корневым **ъ: волк*, acc. *валкá/вóлка; плот*, ńm. *платá*, gen. *плóта/платá*, n.pl. *платá*.

Таким образом, у довольно большого числа слов хотя бы в одном говоре зафиксирован а. т. B с рефлексами форм-энклиноменов в n.-acc.sg. (1) Во всех говорах это представлено у зоонимов, **u*-основ *вол*, *сом*, *дрозд*, к которым примыкает **jo*-основа *хорь*. (2) В средне- и южнорусских говорах а. т. B с рефлексами форм-энклиноменов в корне имеется у **i-* и консонантных основ (*гвоздь*, *локоть*, *ноготь*, *лось*), причем в Нов. Пен. более последовательно, чем в Пуст., тогда как севернорусские говоры Сл. и Ихалицы [Брок 1907] переводят эти слова в а. т. C. (3) **o-* и **u*-основы *плод*, *рог*, *мост*, *нос* встречаются только с корневой /o/ и обнаруживают следы нафлексионного ударения в формах ед. числа в говорах Сл., Пуст., тогда как в Нов., Пен., Их., Леке [Шахматов 1914] у них стабильно представлено начальное ударение. (4) **o-* и **u*-основы *клоп*, *сноп*, *пост*, *под* в среднерусских говорах Пуст. и Леки стабильно показывают а. т. B и корневую /ω/ (рефлекс а. п. *b*), тогда как в севернорусском говоре Сл. представлено колебание а. т. B/C, B/D и корневая /o/, т. е. обнаруживается хотя бы в качестве вариантного, рефлекс а. п. *d*. Говоры Пен., Нов. занимают промежуточное место.

Несомненно, что системы Сл., Пуст., Нов., Пен., Их. и Леки являются развитием одной и той же акцентной системы. Представляется, что архаичное состояние системы представлено в Пуст. и Нов., с рефлексами а. п.*b* в (4) группе примеров и а. п.*d* — в (1) группе примеров (у **u-*, **i-* и консонантных основ). Различие между Пуст. и Нов., помимо акцентуации и огласовки отдельных основ, заключается в последовательном наосновном ударении форм ед. числа **o-*основ а. п. *d* в говоре Нов. и следах нафлексионного ударения — в Пуст. В севернорусском говоре Сл. лишь зоонимы стабильно сохраняют а. т. B < а. п. *d*, тогда как **i-* и консонантные основы а. п. *d* последовательно переведены в а. т. C, а у твердых основ наблюдается взаимное влияние а. т. B и C в формах ед. числа.

4.4. Различия в огласовке корней бесприставочных отглагольных существительных муж. рода, оканчивающихся на **oj* (**jo-*основы, **ju-*основа *зной*), связаны с праславянской а. п. производящего глагола. Те из них, которые имеют формы мн. числа, синхронно относятся к **а. т. D**, остальные имеют накоренное ударение в формах ед. числа.

Производные от глаголов а. п. *a:* Нов. *бωй*, gen. *бώйа*, n.pl. *баи́*, Пуст. *бωй*, gen. *бώйъ*, n.pl. *бои́*; Сл. *бωй/бой;* Нов. *рωй* (но Пуст. *рой*, gen. *ро́йъ*, in. *ро́йъм*, n.pl. *рои́*); Сл. *стрώй*, Пуст. *стрωй*, gen. *стрώйъ*, in. *стрώйъм*.

Производные от глаголов а. п. *c:* Пуст. *вой*, gen. *во́йъ*; Пуст. *гно́й*, gen. *гно́йъ* (но также *гнώй*, gen. *гнώйъ*), Сл. *гной; зной*, gen. *зно́йъ* (всюду).

Литература

1. *Васильев 1929* — Л. Л. Васильев. О значении каморы в некоторых древнерусских памятниках XVI–XVII вв. К вопросу о произношении звука о в великорусском наречии. Л., 1929.
2. *Дыбо 2000* — В. А. Дыбо. Морфонологизированные парадигматические акцентные системы. Типология и генезис. Т. 1. М., 2000.
3. *Дыбо, Замятина, Николаев 1990* — В. А. Дыбо, Г. И. Замятина, С. Л. Николаев. Основы славянской акцентологии. М., 1990.
4. *Зализняк 1985* — А. А. Зализняк. От праславянской акцентуации к русской. М., 1985.
5. *Крылов, Тер-Аванесова 2006* — Крылов С. А., Тер-Аванесова А. В. Лексико-грамматические базы данных как инструмент диалектологического описания // Лауфер Н. И., Нариньяни А. С., Селегей В. П. (ред.), Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог'2006» (Бекасово, 31 мая — 4 июня 2006 г.). М.: Издательство РГГУ, 2006, с. 493–497.
6. *Крылов, Тер-Аванесова 2009* — Крылов С. А., Тер-Аванесова А. В. Использование лексико-грамматических баз данных в русской диалектной лексикографии // Кибрик А.Е. (ред.), Компьютерная лингвистика и интеллектуальные технологии. Вып. 8 (15). По материалам международной конференции «Диалог'2009» (Бекасово, 27–31 мая 2009 г.). М.: РГГУ, 2009, с. 471–475.
7. *Шахматов 1914* — А. А. Шахматов. Описание Лекинского говора Егорьевского уезда Рязанской губернии // Известия ОРЯС 1913. СПб, 1914. Т. 18, кн. 4. С. 171–220.

К определению составного союза (анализ *даже если*)[1]

Compound conjunction vs. word combination (semantics of *dazhe esli* ‘if ever’)

Урысон Е. В. (uryson@gmail.com)

Институт русского языка им. В. В. Виноградова РАН

Проанализирована семантика союза *если*. Показано, что значение сочетания *даже если*, которое в грамматике и типологии считается союзом, выводимо из значения ‘даже’ и ‘если’. Специфика сочетания состоит в том, что союз *если* выступает в нем в своей неосновной модификации; вне сочетания с *даже* союз *если* выступает в ней лишь в ограниченном круге контекстов.

1. Постановка задачи

Предлагаемая работа посвящена семантическому анализу сочетания союза *если* с частицей *даже*, ср. *Даже если погода будет плохая, мы пойдем в лес*.

Принято считать, что сочетание *даже если / если даже* представляет собой отдельный союз [Грамматика-80; Томмола 2004; Храковский 2004]. Однако работ, посвященных обоснованию этого решения, насколько нам известно, не существует. По-видимому, такая трактовка выбрана по интуиции. Мы попытаемся эксплицировать эту исследовательскую интуицию и уточним статус *даже если / если даже* в лексической системе языка.

Составной союз «представляет собой соединение двух или более элементов, каждый из которых одновременно существует в языке и как отдельное слово» [Грамматика-80: 716]. Данное определение позволяет выделить составные союзы из множества союзов вообще, однако не позволяет отличить составной союз от сочетания служебных слов. В частности, остается неясным, что представляет собой цепочка *даже если*: возможно, это составной союз, однако это может быть и просто сочетание союза *если* и частицы *даже*.

Во многих случаях отличить последовательность служебных слов от составного союза позволяют синтаксические критерии. Например, союз *то… то* внешне совпадает с последовательностью местоимений *то…то*. Ср. *То дождь, то ветер* [союз] VS. *То не ветер ветку клонит, / Не дубравушка шумит, / То мое сердечко стонет* [два местоимения] (Н. Стромилов). Однако синтаксические функции союза *то…то* резко отличаются от синтаксических функций местоимения *то*, поэтому проблем при определении статуса последовательности *то…то*, вообще говоря, не возникает.

В случае *даже если* синтаксические критерии не работают, так как последовательность *даже если / если даже* синтаксически ведет себя как союз *если*.

Очевидно, что составной союз отличается от простой последовательности отдельных слов, прежде всего, с точки зрения семантики: значение составного союза не сводимо к «сумме» значений его элементов. Следовательно, для того чтобы отнести некоторую последовательность служебных слов к составным союзам, необходим семантический анализ этой последовательности.

Логика нашего рассуждения такова. Допустим, что *даже если / если даже* — это просто сочетание двух служебных слов. Тогда семантика этого сочетания представляет собой «сумму» значений входящих в него компонентов ‘если’ и ‘даже’, без каких-либо нестандартных наращений или изменений этих смыслов. Если же наше сочетание — это действительно отдельный союз, то его семантика не может быть полностью извлечена из значений составляющих частей ‘если’ и ‘даже’. Если данное сочетание обладает определенными синтаксическими свойствами, то его можно трактовать как союз.

[1] Работа выполнена при финансовой поддержке Программы фундаментальных исследований ОИФН РАН «Генезис и взаимодействие социальных, культурных и языковых общностей» и гранта НШ-3205.2008.6. Автор благодарит анонимного рецензента за ценные критические замечания.

Следовательно, для ответа на поставленный выше вопрос нужно подвергнуть сочетание *даже если / если даже* композициональному анализу и выявить вклад в семантику всего сочетания его составных частей. (Вариативный порядок компонентов *даже если* vs. *если даже* мы здесь не описываем.)

Очертим основные проблемы, возникающие при семантическом анализе высказываний с *даже если / если даже*.

Хорошо известно, что частица *даже* выражает некоторое модальное значение, которое определенным образом взаимодействует со значением остальной части высказывания. Возьмем, например, следующие примеры:

(1) *Петя пошел в кино*.

(2) *Даже Петя пошел в кино*.

Семантическую структуру примера (2) можно представить так (А. Вежбицкая; цит. по книге [Апресян 1974: 33, 68]):

(2а) (i) 'Петя пошел в кино';

(ii1) 'другие люди пошли в кино';

(ii2) 'говорящий ожидал, что Петя не пошел в кино'.

Более точно [Богуславский 1996; см. также: Крейдлин 1975]:

(2б) (i) 'Петя пошел в кино';

(ii1) 'другие люди пошли в кино';

(ii2) 'можно было с бóльшим основанием ожидать, что другие люди пошли в кино, чем то, что Петя пошел в кино'.

Компонент (i) этих структур — это, очевидно, целиком значение примера (1). Компоненты (ii1–2) представляет собой значение частицы *даже*. В обычном случае семантическая структура высказывания *S*, содержащего частицу *даже*, может быть представлена как состоящая из двух частей: 'S ´' и 'даже', где *S ´* — это высказывание *S* за вычетом *даже*. Иными словами, высказывание типа (2), с частицей *даже*, описывает ту же ситуацию, что и соответствующее высказывание без *даже*, ср. (1), отличаясь от него лишь выражением модальности. Естественно предположить, что частицу *даже* можно ввести практически в любой правильный пример и полученное высказывание будет отличаться от исходного только модальным компонентом.

Возьмем теперь какое-нибудь стандартный пример с *если* и введем в него частицу *даже*. Ср.

(3) *Я останусь у них, если они меня пригласят*.

(4) ??*Я останусь у них, даже если они меня пригласят*.

Соотношение примеров внутри этой пары совсем не такое, как между примерами (1) и (2). Действительно, высказывание (4) семантически или по крайней мере прагматически аномально. Семантически и прагматически нормальным будет отрицание высказывания (4), ср.

(4а) *Я останусь у них, даже если они меня не пригласят*.

Ср. аналогичные примеры:

(5) *Если погода будет хорошая, мы пойдем купаться*.

(6) ??*Даже если погода будет хорошая, мы пойдем купаться*.

(6а) *Даже если <если даже> погода будет хорошая, мы купаться не пойдем*.

Высказывание (6) и в этой группе примеров как минимум прагматически аномально. Нормально высказывание (6а), с отрицанием в главном предложении.

(7) *Если ты будешь работать лучше всех, тебя повысят*.

(8) ??*Даже если <если даже> ты будешь работать лучше всех, тебя повысят*.

(8а) *Даже если <если даже> ты будешь работать лучше всех, тебя не повысят*.

Пример (8), отличающийся от высказывания (7) только наличием *даже*, аномален, по крайней мере прагматически. Частица *даже* и здесь требует отрицания в главном предложении, ср. (8а).

Можно предположить, что сочетание *даже если / если даже* отрицательно поляризовано, т. е. тяготеет к отрицательным контекстам. И союз *если*, и частица *даже* в этом отношении нейтральны. По-видимому, отрицательная поляризованность *даже если / если даже* и наводит на мысль, что это отдельный союз.

Однако в некоторых случаях *даже если / если даже* не требует отрицательного контекста. Ср.

(9) *Если Петя ее обидит, она его простит*.

(10) *Даже если <если даже> Петя ее обидит, она его простит*.

Соотношение между примерами в этой паре абсолютно стандартно — оно ничем не отличается от соотношения между (1) и (2). Семантическая структура высказывания (10) состоит из смысла высказывания (9) и модального компонента, выражаемого *даже*.

Наконец, существует еще третья группа контекстов, в которых *даже если / если даже* как будто не поляризовано отрицательно, но при этом как будто и не сводимо к «сумме» ‘если’ и ‘даже’. Ср.

(11) *Если Петя будет работать лучше всех, его уволят*.

(12) *Даже если <если даже> Петя будет работать лучше всех, его уволят*.

Примеры (11) и (12) выражают разные смыслы, и различие между ними не сводится к выражению модальности. В (11) подразумевается, что сотрудник, в частности Петя, не должен выделяться — нужно быть как все, иначе уволят. В (12), напротив, подразумевается, что работать лучше всех — это хорошо, это та причина, по которой человека обычно ценят и не увольняют.

Требуется выяснить, за счет чего возникает то или иное понимание высказывания с *даже если / если даже*: потому ли, что перед нами разные значения союза (или сочетания) *даже если / если даже*, или по каким-то другим, более тонким причинам.

Мы надеемся получить ответ на поставленные выше вопросы, подвергнув *даже если / если даже* композициональному семантическому анализу.

Частица *даже* подробно описана в работах [Крейдлин 1975; Богуславский 1985; 1996]. Поэтому в нашу задачу входит, прежде всего, анализ союза *если*.

2. Семантический анализ союза *если*

Продемонстрируем, что союз *если*, хотя и является классическим примером семантического примитива [Жолковский 1964; Мельчук 1974; Вежбицкая 1996а,1996б; Wierzbicka 1997], однако разложим на четкие семантические компоненты (см. также [Урысон 2001; 2010]).

2.1. Структура многозначности союза *если*

А. Центральное значение союза *если* — лексема «*если* гипотезы»[2]:

(13) *Если будут пробки (P), то мы опоздаем на самолет (Q)*;

(14) *Мы пойдем купаться (Q), если будет хорошая погода (P)*;

(15) *Если Петя приехал во вторник (P), он уже все узнал от Андрея (Q)*.

Очевидно, что союз *если* в приведенных примерах — это средство, с помощью которого мы сообщаем, что и P, и Q — это всего лишь наше предположение, наша гипотеза (ср. [Ducrot 1972]), а не описание реального или даже вымышленного положения дел.

Б. Лексема «*если* обобщения»:

(16) *Если были пробки (P), то мы опаздывали (Q)*;

(17) *Если шел дождь (P), лепили на веранде из пластилина (Q)* (С. Довлатов);

(18) *Он счастлив (Q), если ей накинет / Боа пушистый на плечо (P)* (А. С. Пушкин).

Союз «*если* обобщения» сближается по значению с союзом *когда*. Ср.

(16а) *Когда были пробки (P), мы опаздывали (Q)*;

(17а) *Когда шел дождь (P), лепили на веранде из пластилина (Q)*.

В работах [Урысон 2001; 2010] показано, что «*если* обобщения» — это контекстно-обусловленная модификация лексемы «*если* гипотезы».

В. Лексема «*если* данного положения дел»:

(19) *Если уж мой лучший друг решил остаться (P), я остаюсь тоже (Q)*.

(20) *— У него там мать замдекана.*
— Ну, если у него мать замдекана (P), он, конечно, будет учиться бесплатно (Q).

(21) *— Я получил пять!*
— Ну, если у тебя по философии пятерка (P), то теперь тебе бояться нечего (Q).

Союз «*если* данного положения дел» синонимичен союзу *раз*. Ср. (19)–(21) и следующие примеры с *раз*:

(19а) *Раз уж мой лучший друг решил остаться (P), я остаюсь тоже (Q)*.

[2] В соответствии со словоупотреблением, принятым в Московской семантической школе, слово, взятое в отдельном значении, называется лексемой. Многозначное слово представляет собой тогда упорядоченный набор лексем. Описать многозначное слово — значит описать по возможности все его лексемы, указав семантические связи (мосты) между ними.

(20а) *— У него там мать замдекана.*
— Ну, раз у него мать замдекана (P), он, конечно, будет учиться бесплатно (Q).

(21а) *— Я получил пять!*
— Ну, раз у тебя по философии пятерка (P), то теперь тебе бояться нечего (Q).

Многие фразы с *если* омонимичны: они могут описывать как гипотезу говорящего, так и данное положение дел. Ср.

(22) *Если Петя заболел, она не придет* [говорящий либо предполагает, что Петя заболел, либо знает это].

Эта омонимия снимается широким контекстом, ср.

(22а) *— Оказывается, Петя заболел!*
— Ну, если Петя заболел, она не придет [лексема «*если* данного положения дел»].

(22б) *Не знаю, как они. Если Петя заболел, она не придет* [лексема «*если* гипотезы»].

Наличие таких неоднозначных фраз — свидетельство того, что союз «*если* данного положения дел» и союз «*если* гипотезы» — это две разные лексемы *если*.

Г. Лексемы «*если* умозаключения».

(23а) *Если они успели на последнюю электричку (P), они уже дома (Q).*

(23б) *Если они уже дома (P), они успели на последнюю электричку (Q).*

Для простоты ограничимся одним — гипотетическим — пониманием этих фраз. В обоих примерах ситуация ‘они успели на последнюю электричку’ является тем фактором, тем условием, благодаря которому имеет место ситуация ‘они уже дома’. Фраза (23а) подает это самым естественным образом: условие ‘они успели на последнюю электричку’ вводится союзом *если* и оформляется как придаточное предложение, а следствие ‘они уже дома’ оформляется главным предложением. Но в примере (23б) условие и следствие поменялись местами: условие представлено в главном предложении, а союз *если* вводит следствие. Очевидно, что в примере (23б) союз *если* не только описывает связь между двумя ситуациями, но еще и специально указывает на ход мысли говорящего, на его умозаключение. Будем говорить, что в примерах типа (23б) представлен союз «*если* гипотезы + умозаключения». Это квазиконверсив союза «*если* гипотезы». Приведем аналогичные примеры:

(24а) *Если они не спят (P), у них в окнах горит свет (Q)* [ситуация ‘они не спят’ — фактор, благодаря которому существует ситуация ‘у них в окнах горит свет’].

(24б) *Если у них в окнах горит свет (P), они не спят (Q)* [«объективный» смысл высказывания — тот же, что и в примере (24а); однако ситуация ‘у них в окнах горит свет’ является здесь фактором, благодаря которому говорящий делает умозаключение о том, что люди в доме не спят].

Указание на умозаключение может сочетаться и с семантикой «*если* обобщения», и с семантикой «*если* данного положения дел». Перечисление этих случаев выходит за рамки доклада.

Д. Другие контексты с *если*:

(Д1) *Если захочешь есть (P), в холодильнике есть мясо (Q)* (Дж. Остин); *Если захочешь есть (P), возьми мясо (Q)*; *Если тебе так холодно (P), почему ты сидишь в одной рубашке(Q)?*; *Если хочешь знать правду (P), отец уехал насовсем (Q).*

(Д2) *Если верить газетам (P), все было совсем по-другому (Q)*; *Это, если хотите (P), совершенно особый случай (Q); Если помните (P), мы к вам в пятницу заходили (Q).*

(Д3) *В его глазах он заметил если не страх, то во всяком случае растерянность.*

(Д4) *Если кочевые народы занимались скотоводством (P), то оседлые рано достигли высокой культуры земледелия(Q).*

Контексты типа (А)–(Г) допускают замену *если* на *даже если / если даже*. Требуется описать значение соответствующих лексем союза *если*. Рамки статьи позволяют нам подробно описать лишь лексему «*если* гипотезы». Однако полученные результаты легко распространяются и на другие лексемы *если*.

2.2. Центральное значение союза *если*: «*если* гипотезы»

(13) *Если будут пробки (P), то мы опоздаем на самолет (Q).*

(14) *Мы пойдем купаться (Q), если будет хорошая погода (P)*;

(15) *Если Петя приехал во вторник (P), он уже все узнал от Андрея (Q).*

Союз «*если* гипотезы» указывает на то, что и P, и Q — это всего лишь наше предположение, наша гипотеза. Именно поэтому во многих языках союз со значением ‘если’ требует постановки предиката вводимой им пропозиции в специальное — условное — наклонение. В каких случаях мы строим такие гипотезы?

Ясно, что говорящий прежде всего не знает, какая ситуация — P или не-P — имеет место в описываемый отрезок времени. Кроме того, говорящий считает, что в этот отрезок времени ситуация P возможна. При этом, строя гипотезу ‘P имеет место’, говорящий безусловно отдает себе отчет в том, что может иметь место как P, так и не-P [Грамматика-80; Падучева 1985; 2004:103; Ляпон 1986; Вежбицкая 1996б; Санников 2001/2008]. Однако гипотеза говорящего не ограничивается содержанием ‘P имеет место’. Говорящий представляет, как развиваются события в рамках его гипотезы, а именно, какая еще ситуация Q, связанная с P, имеет тогда место.

Сказанное можно записать в виде следующего выражения:

(I) *Если P, то Q [Если будут пробки (P), мы опоздаем на самолет (Q)]* ≈

(Ia) ‘говорящий не знает, какая ситуация имеет место в описываемый отрезок времени’;

(Ib) ‘говорящий считает, что в этот отрезок времени возможна [= может иметь место] ситуация P, возможна [= может иметь место] ситуация не-P’;

(Ic) ‘говорящий представляет [≈ строит гипотезу]: имеет место ситуация P’;

(Id) ‘говорящий утверждает, что в рамках этой гипотезы события развиваются так: имеет место ситуация Q, связанная с ситуацией P’.

В ирреальных контекстах экспликация (I) модифицируется [Урысон 2001; 2010].

Статус выражения (I) будет рассмотрен ниже. Сейчас попытаемся ответить на более частный вопрос: каким образом могут быть связаны ситуации P и Q.

Оказывается, что союз *если* служит для обозначения целого спектра возможных видов связи между двумя ситуациями.

Начнем с логически самой простой связи.

(i) ситуация P является причиной существования ситуации Q

(13) *Если будут пробки (P), мы опоздаем на самолет (Q)* [пробки (P) — причина опоздания на самолет (Q)].

(25) *Если мы сдадим работу в срок (P), премия нам обеспечена (Q)* [выполнение работы вовремя (P) — причина получения премии (Q)];

(26) *Если они не достали билетов на автобус (P), им придется ехать поездом (Q)* [отсутствие билетов на автобус (P) — причина того, что им придется ехать поездом (Q)].

В ряде работ считается, что союз *если* выражает только данное отношение [Comrie 1986; Падучева 2004: 103]. Однако две ситуации могут быть связаны и несколько иначе.

(ii) ситуация P является условием существования ситуации Q

(14) *Мы пойдем купаться (Q), если будет хорошая погода (P)*.

В данном случае ситуация P не создает ситуацию Q, как в примерах выше, а лишь благоприятствует ей, является условием ее существования.

(27) *Если они попадут в Париж (P), то сразу отправятся в Лувр (Q)* [пребывание в Париже (P) — условие посещения Лувра (Q)].

(28) *Если Петя приехал во вторник (P), он уже все узнал от Андрея (Q)* [приезд Пети во вторник — не причина, а обстоятельство, благодаря которому он все знает от Андрея].

Во многих случаях трудно определить, какая именно связь — причинная или условная — выражается в предложении вида *Если P, то Q*. Ср.

(29) *Если ты не проспишь (P), ты сможешь увидеть рассвет (Q)*.

Причина и условие не всегда различимы по природе самих вещей. Отметим, что эти языковые концепты лежат в основе философских понятий причины, условия и причинности, однако и с точки зрения философии «различие между причиной и условием относительно» [ФЭС, статья «Причина и следствие»].

(iii) ситуация P влияет на положение дел и тем самым обуславливает ситуацию Q

(30) *Если дети пойдут в поход (P), они обязательно возьмут с собой гитару (Q)*.

(31) *Если у них родится мальчик (P), они назовут его Иваном (Q)* [пример из работы В. З. Санникова [2001/2008]].

(32) *Если к нам приедут гости (P), мы отведем им самую лучшую комнату (Q)*.

Вряд ли ситуация P в этих высказываниях является причиной или условием существования ситуации Q: такая трактовка кажется натянутой [Вежбицкая 1996б; Wierzbicka 1997]. Тем не менее, и в этих контекстах между P и Q имеется некая, хотя и трудно эксплицируемая, каузальная зависимость. Иногда эта зависимость похожа на причинную: так, в (31)–(32) ситуация Q не может существовать без ситуации P. Иногда эта зависимость напоминает скорее условную, ср. (30). Но ни (30), ни (31)–(32) не допускают соответствующих перифраз. Абсурдно высказывание **Дети обязательно возьмут с собой гитару при условии, что они пойдут в поход*. Бессмысленны высказывания: **Мы отвели гостям самую лучшую комнату, потому что они к нам приехали*; **Они назвали мальчика Иваном, потому что он у них родился* (равно как и абсурдны высказывания **Они назвали ребенка Иваном, потому что он мальчик <потому что у них родился мальчик>*). Безусловно, в подобных случаях речь идет о влиянии ситуации P на некоторое общее положение дел, и существование ситуация Q — одно из следствий этого влияния. Однако описать это влияние конкретнее, по-видимому, невозможно. Существенно, что данный тип каузальной связи между P и Q не конкретизирован в самом естественном языке. Подобную каузальную зависимость можно назвать каузальностью, или обусловленностью, в широком смысле. О понятии обусловленность см. работу [Евтюхин 1996].

(iv) ситуации P и Q вызваны одной и той же причиной или существуют благодаря одному и тому же условию

(33) *Если американцы первыми полетят на Марс (P), они и на Венеру полетят первыми (Q)*.

(34) *Если он выиграет первенство мира (P), он и на Олимпиаде станет победителем (Q)*.

Очевидно, что, например, ситуация ‘американцы первыми полетят на Марс’ не может быть ни причиной, ни условием того, что они первыми полетят и на Венеру. Аналогичным образом устроена и фраза (34). Ситуации P и Q связаны в обоих случаях опосредованно — они существуют благодаря одной и той же причине или возникают при одних и тех же условиях: применительно к (33) это высокий уровень развития науки и техники, в случае (34) это мастерство данного спортсмена.

Похожим образом устроены и следующие высказывания, предполагающие ситуацию гадания:

(35) *Если сумма цифр будет четной (P), он меня любит (Q)*;

(36) *Если сейчас из-за угла выедет машина (P), я сдам экзамен (Q)*.

Прототипически гадание основано на следующем представлении об устройстве мира. «В мире нет случайностей. Одна и та же сила устраивает все ситуации. Человек не знает, как действует эта сила, поэтому он не знает будущего. Однако даже самая незначительная ситуация P — это проявление действия данной силы. Эта сила может подать человеку знак, по которому можно узнать, как она действует в очень важном случае, в частности, можно узнать, будет иметь место ситуация Q или не-Q». Поскольку одна и та же сила устраивает все ситуации, то значит, все ситуации имеют какую-то общую, универсальную причину. Тем самым, фразы (35)–(36) предполагают опосредованную общую причину у двух, казалось бы, совершенно независимых ситуаций. При этом ситуация Q важна для говорящего, а ситуация P мыслится как знак ее существования (благодарим за это замечание Р. И. Розину). Очевидно, что этот знак «подается» той же силой, от которой зависит и существование ситуации Q. (Представления, которые лежат в основе гадания, уходят из обыденного сознания людей, сохраняясь при этом в стереотипах речи. Подобные ситуации хорошо известны лингвистам [Абаев 1934/1995].)

В случае обусловленности двух ситуаций P и Q каким-то третьим фактором, союз *если* уже не вводит ситуацию-каузатор (она остается невыраженной) — он просто указывает на сосуществование, сопутствование двух ситуаций. Благодаря этому в данном классе высказываний союз *если* приобретает определенное сходство с симметричными предикатами. В частности, для данных высказываний справедливо следующее квазисинонимическое преобразование:

(37) *Если A, то B ≈ Если B, то A*.

Примеры:

(38) *Если сумма цифр будет четной (A), он меня любит (B) ≈ Если он меня любит (B), сумма цифр будет четной (A)*;

(39) *Если сейчас из-за угла выедет машина (A), мне достанется второй билет (B) ≈ Если мне достанется второй билет (B), то сейчас из-за угла выедет машина (A)*.

Правда, в некоторых случаях преобразование (37) влечет довольно существенное изменение смысла: меняется последовательность описываемых ситуаций. Ср.

(40) *Если американцы первыми полетят на Марс (A), они и на Венеру полетят первыми (B)* [полет на Марс раньше полета

на Венеру] ≈ *Если американцы первыми полетят на Венеру (B), они и на Марс полетят первыми (A)* [полет на Венеру раньше полета на Марс].

Однако сама по себе опосредованная связь между описываемыми ситуациями сохраняется и в этом случае. Между тем в случае непосредственной каузальной связи между P и Q преобразование (52) абсолютно недопустимо. Ср. нормальное высказывание (14) и абсурдное (14a):

(14) *Мы пойдем купаться (A), если будет хорошая погода (B)*.

(14a) *Если мы пойдем купаться (A), будет хорошая погода (B)*.

Аналогичным образом нормально высказывание (13) и абсурдно (13a):

(13) *Если будут пробки (A), то мы опоздаем на самолет (B)*.

(13a) *Если мы опоздаем на самолет (B), будут пробки (A)*.

Таково кардинальное различие между данным типом связи между двумя ситуациями и непосредственной каузальной зависимостью между ними.

2.3. Промежуточный итог

Попытаемся обобщить, что же выражает союз *если* в перечисленных случаях (i)–(iv). Для этого нам понадобятся два понятия — ситуация и положение дел.

Ситуация — это то, что обозначается предикатом со всеми его зависимыми. В частности, простое предложение *P* обозначает ситуацию ‘P’. Именно так мы употребляли термин *ситуация* выше, в выражениях типа ‘говорящий представляет: имеет место ситуация P’.

Под *положением дел* будем понимать совокупность ситуаций, как-то связанных друг с другом. Так, текст: *Начался дождь (P). Все побежали к метро (Q), и улица быстро опустела (R)* описывает положение дел, состоящее из трех ситуаций — P, Q и R. Аналогичным образом, высказывание *Когда идет снег (P), я всегда вспоминаю Москву (Q)* описывает положение дел, состоящее из ситуаций P и Q.

В контекстах (i)–(iv) речь идет о том, что некоторая ситуация влияет на имеющееся положение дел. Эта ситуация может быть не названа — случай (iv), ср. *Если американцы первыми полетят на Марс (P), они и на Венеру полетят первыми (Q)*; *Если сумма цифр будет четной (P), он меня любит (Q)*. Здесь и главное, и придаточное предложения описывают результат влияния на положение дел некоторой третьей, неназванной ситуации.

Однако в контекстах (i)–(iii) обозначены и ситуация, влияющая на положение дел, и ситуация, являющаяся следствием этого влияния. При этом информация о том, какая ситуация является фактором, а какая — следствием этого фактора, входит в значение союза *если*. Она выражается в синтаксическом оформлении пропозиций: пропозиция, обозначающая причину или условие, всегда вводится *если* и составляет придаточное, а пропозиция, обозначающая результат действия данной причины (или условия), всегда оформляется как главное предложение. Поэтому нормальна фраза типа *Если будет хорошая погода, мы устроим маленький поход*, но аномально (абсурдно, бессмысленно) высказывание *Если мы устроим маленький поход, будет хорошая погода*. Второе высказывание может быть понято только в том смысле, что наш поход повлияет на погоду, а это противоречит общим знаниям о мире.

Из всего сказанного ясно, что в контекстах типа (i)–(iv) слово *если* ведет себя как каузальный союз с весьма широким значением. При этом характер связей между ситуациями P и Q может быть описан через вполне простые понятия. Поэтому мы можем конкретизировать компонент (Iг) выражения (I), описывающего значение союза *если*. Для удобства выпишем всю дополненную экспликацию союза *если*.

(II) *Если P, то Q [Если будут пробки (P), мы опоздаем на самолет (Q)] ≈*

(IIa) ‘говорящий не знает, какая ситуация имеет место в описываемый отрезок времени’;

(IIb) ‘говорящий считает, что в этот отрезок времени возможна [= может иметь место] ситуация P, возможна [= может иметь место] ситуация не-P’;

(IIc) ‘говорящий представляет [≈ строит гипотезу]: имеет место ситуация P’;

(IId) ‘говорящий утверждает, что в рамках этой гипотезы имеет место следующее:

(IId-1) ситуация P является причиной или условием существования ситуации Q; поэтому имеет место ситуация Q;

или

(IId-2) ситуация P, не являясь причиной или условием существования ситуации Q, влияет на имеющееся положение дел; в результате имеет место ситуация Q;

или

(IId-3) существует некоторая ситуация n, влияющая на имеющееся положение дел; в результате имеют место ситуации P и Q; они вызваны одной и той же причиной n или возникают в результате одного и того же условия n'.

Пока союз *если* как будто не представлял особых трудностей для описания. Однако специфика этого союза состоит в том, что он употребляется еще в одном типе контекстов, которые в каком-то смысле противоположны контекстам, описанным выше.

(v) ситуация P не влияет на положение дел

(41) *Если он сделает что-нибудь не так (P), мама его простит (Q)*.

(42) *Если Коля опоздает (P), его не оштрафуют (Q)*.

(43) *Если Петя ее обидит (P), она будет продолжать улыбаться (Q)*.

(44) *Если они окажутся в Париже (P), они в музеи ходить не будут (Q) — их интересуют только магазины*.

(45) *Если я заболею, к врачам обращаться не стану* (Я. Смеляков) [Санников 2001/2008: 418].

На подобные высказывания обратила внимание А. Вежбицкая, см. [Вежбицкая 1996б; Wierzbicka 1997] — именно они препятствуют тому, чтобы считать союз *если* каузальным, и именно они делают этот союз нетолкуемым. Подобные контексты четко демонстрируют, что семантика *если* не сводится к обозначению условия, или причины, или какой-либо каузальной зависимости. Эти высказывания представляют для нас особый интерес.

Данный тип контекстов как будто похож на контексты (i)–(ii), но в отличие от них, здесь речь идет о том, что ситуация P никак не влияет на положение дел. Получается, что союз *если* может оформлять противоположные типы отношений между P и Q. Ср.

(41а) *Если он сделает что-нибудь не так (P), мама его не простит (Q) — Если он сделает что-нибудь не так (P), мама его простит (Q)*.

(42а) *Если Коля опоздает (P), его оштрафуют (Q) — Если Коля опоздает (P), его не оштрафуют (Q)*;

(43а) *Если Петя ее обидит (P), она перестанет улыбаться (Q) — Если Петя ее обидит (P), она будет продолжать улыбаться (Q)*;

(44а) *Если они окажутся в Париже (P), они будут ходить по музеям (Q) — Если они окажутся в Париже (P), они не будут ходить по музеям (Q)*;

Союз *если* выражает здесь особый смысл. Существенно, однако, что это возможно лишь в определенном типе контекстов. Требуется, чтобы ситуация P принадлежала к набору всем известных, житейских ситуаций, которые обычно или часто влияют на имеющееся положение дел, но могут и не повлиять на него. Таковы ситуации опоздания (могут заметить и оштрафовать, а могут и не заметить и не оштрафовать), супружеской измены (одни не прощают, а другие прощают), обиды (кто-то реагирует на обиду, а кто-то — нет) и т. п. Набор таких ситуаций, по-видимому, невелик. Если ситуация P не принадлежит к этому набору, то союз *если* всегда выражает влияние P на положение дел. Ср.

(46) *Если бумага будет черной (P), буквы на ней проступят (Q) — Если бумага будет черной (P), буквы на ней не проступят (Q)* [ситуация P в обоих случаях влияет на положение дел и порождает ситуацию Q].

Итак, союз *если* выражает особый, некаузальный смысл лишь в определенном круге контекстов. Но тогда перед нами не особое значение союза *если*, а лишь его контекстная модификация. Действительно, фразы типа *Если Коля опоздает (P), его оштрафуют (Q) — Если Коля опоздает (P), его не оштрафуют (Q)* различаются на отрицание 'не', и именно оно, а не союз *если*, создает «противоположное» понимание данных высказываний.

Контексты типа (v) близки высказываниям с уступительным союзом *хотя* в его центральном значении [Грамматика-80]. Ср. примеры внутри следующих пар:

(47) а. *Если он сделает что-нибудь не так (P), мама его простит (Q)* — б. *Хотя он сделает что-нибудь не так (P), мама его простит (Q)*.

(48) а. *Если Коля опоздает (P), его не оштрафуют (Q)* — б. *Хотя Коля опоздает (P), его не оштрафуют (Q)*.

Все же между *если* и *хотя* существует важное семантическое различие. Сравним фразы (47а) и (47б). Одно из различий между ними очевидно: в высказывании с *хотя* утверждается, что субъект сделает что-нибудь не так, а в высказывании с *если* та же пропозиция P выражает всего лишь предположение говорящего. Это различие целиком обусловлено описанной выше семантикой союза «*если* гипотезы». Но между данными фразами есть и другое, более тонкое различие. Оно состоит в следующем.

Союз *хотя* в (47а) указывает на то, что ситуация P ‘он сделает что-нибудь не так’ — это обычная, нормальная причина наказания субъекта, т. е. ситуации не-Q. Ситуация Q ‘мама его простит’ в данном случае имеет место вопреки обычному, ожидаемому положению вещей. Информация о нарушении обычного порядка вещей выражается союзом *хотя*, а сами «законы мира» известны людям из жизни — это так называемая «наивная энциклопедия». Что касается союза *если* во фразе (47б), то сам по себе он не содержит никакого подобного указания. Скорее наоборот, в высказывании (47б) указываются две возможности: ‘простят’ — ‘не простят’, причем ситуация ‘мама его простит’ воспринимается как более нормальная, естественная, а не как нарушение «житейских закономерностей». Аналогичным образом обстоит дело и в примерах (48а)–(48б).

Попытаемся понять, что выражает союз *если* в подобных контекстах.

Ситуация P в случае (v) мыслится как некая возможная причина или возможное условие другой ситуации: опоздание — обычно причина штрафа, обида лишает человека возможности улыбаться и т. п. Суть в том, что эта причина (или условие) остаются потенциальными: ситуация P не влияет на положение дел. Тем не менее ситуация P, подобно «настоящей» причине, всегда вводится союзом *если* — иной порядок пропозиций радикально меняет смысл высказывания или же приводит к почти абсурдному результату. Ср. *Если он ей опять изменит, она его простит* [измена не влияет на положение дел] VS *Если она его простит, он ей опять изменит* [полученное прощение влияет на положение дел, так что человек опять изменяет]; *Если Петя опоздает, его не оштрафуют* [опоздание не влияет на положение дел] VS *Если Петю не оштрафуют, он опоздает* [прагматически странное высказывание, с трудом поддающееся осмыслению].

Быть может, союз *если* указывает на то, что ситуация P не влияет на положение дел? Мы вынуждены признать, что в семантику *если* это указание не входит — соответствующая информация передается контекстом. Что же остается на долю союза?

В примерах типа *Если он опоздает, его не оштрафуют*, безусловно, обозначена какая-то связь между P и Q, причем эта связь как будто имеет отношение к каузальной зависимости, но при этом она таковой не является. Перед нами какая-то самая естественная, понятная связь между ситуациями: P — это некоторая «исходная» ситуация, отправная точка нашей гипотезы, а Q — развитие гипотезы, то положение дел, которое имеет место в ее рамках. Язык использует для обозначения такой связи между ситуациями союз *Если P, то Q*. Мы, однако, не находим в языке никакого знаменательного, т. е. не служебного слова для обозначения этой связи. Тем самым, данная связь между ситуациями не поддается экспликации — она может быть обозначена только самим союзом *если*. Эта связь отчасти напоминает каузальную, но ее невозможно описать в привычных терминах каузальной зависимости. Мы можем дать ей лишь условное название, например такое: «псевдозависимость ситуации Q от ситуации P». Но разумеется, это только ярлык, а не экспликация семантики союза *если* в данном типе высказываний. Мы не можем истолковать союз *если* в контекстах типа (v) через более простые понятия.

Вернемся к нашей попытке толкования союза «*если* гипотезы» — к выражению (II). Для того чтобы учесть контексты (v), это выражение нужно дополнить. Требуется ввести в список связей между ситуациями P и Q указание на тип связи между ситуациями в этих контекстах. Обсуждаемый компонент мог бы иметь такой вид: ‘имеет место псевдозависимость ситуации Q от ситуации P’.

Ясно, однако, что данное выражение не может быть компонентом аналитического толкования: обсуждаемый тип связи между ситуациями не определяется, а просто обозначается некоторым условным ярлыком. В семантике союза *если* выделяется специфический компонент, для выражения которого мы не находим в естественном языке никакого подходящего слова. Перед нами «довербальный семантический элемент». Слово «довербальный» не означает, что данный элемент исторически предшествует обычным семантическим компонентам (так же как слово *недоразвитый* не обозначает начальной или промежуточной стадии развития, после которой наступает норма). Определение «довербальный» указывает на специфическую природу данного элемента, на его особый статус в семантическом метаязыке: в естественном языке нет слова или морфемы для его выражения. Этим данный элемент принципиально отличается от тех элементов, которые обозначаются словами естественного языка.

Исходя из всего сказанного, предлагаем следующую экспликацию союза «*если* гипотезы».

(III) *Если P, то Q [Если будут пробки (P), мы опоздаем на самолет (Q)]* ≈

(IIIa) ‘говорящий не знает, какая ситуация имеет место в описываемый отрезок времени’;

(IIIb) ‘говорящий считает, что в этот отрезок времени возможна [= может иметь место] ситуация P, возможна [= может иметь место] ситуация не-P’;

(IIIc) ‘говорящий представляет [≈ строит гипотезу]: имеет место ситуация P’;

(IIId) ‘говорящий представляет, что в рамках этой гипотезы имеет место следующее:

(IIId-1) ситуация P является причиной или условием существования ситуации Q; поэтому имеет место ситуация Q;

или

(IIId-2) ситуация P, не являясь причиной или условием существования ситуации Q, влияет на имеющееся положение дел; в результате имеет место ситуация Q;

или

(IIId-3) существует некоторая ситуация n, влияющая на имеющееся положение дел; в результате имеют место ситуации P и Q; они вызваны одной и той же причиной n или возникают в результате одного и того же условия n';

или

(IIId-4) имеет место ситуация Q'; "псевдозависимость ситуации Q от ситуации P".

Мы намеренно заключаем последний компонент в обычные, немарровские, кавычки: этим подчеркивается его особая природа — перед нами не компонент аналитического толкования, а довербальный, долексемный семантический элемент, которому мы можем дать лишь условный ярлык.

Поскольку выражение (III) содержит метку некоторого смысла вместо его ясной экспликации, то данное выражение не может претендовать на статус толкования — это всего лишь некая дескрипция союза *если*. Добавим, что в этом выражении фигурирует слово *гипотеза* и выражения *строить гипотезу, в рамках этой гипотезы*, которые семантически несомненно богаче служебного слова *если*, а кроме того стилистически отмечены как книжные. Мы, однако, не можем заменить в предлагаемой дескрипции эти слова более простыми. Тем самым, наше описание подтверждает вывод А. Вежбицкой: «смысл *если* не может быть получен <…> из более простых понятий» [Вежбицкая 1996а: 295]. Но при этом мы выделяем в составе *если* вполне четкие компоненты, из которых только один действительно не поддается экспликации.

Выделение подобного довербального элемента в семантическом разложении слова ставит ряд теоретических проблем, которые подробно обсуждаются в работах [Урысон 2001; 2010].

3. Композициональный анализ *даже если / если даже*

Для простоты изложения разделим контексты с союзом «*если* гипотезы» на два класса.

Первый класс — это каузальные контексты, ср. *Если будут пробки, мы опоздаем на самолет; Мы пойдем купаться, если будет хорошая погода; Если к нам приедут гости (P), мы отведем им самую лучшую комнату (Q); Если американцы первыми полетят на Марс (A), они и на Венеру полетят первыми (B)*. В этих контекстах союз *если* указывает на каузальную зависимость между ситуациями. В них реализуются компоненты (IIId-1)–(IIId3) выражения (III).

Второй класс — это некаузальные контексты. Ср. *Если он ее обидит, она его простит; Если Петя опоздает, его не оштрафуют* и т. п. Здесь реализуется некаузальная модификация союза «*если* гипотезы», содержащая довербальный компонент значения союза, условно названный «псевдозависимость» одной ситуации от другой.

Начнем наше рассуждение с некаузальных контекстов. Будем заменять в этих контекстах союз *если* на *даже если / если даже*. Ср.

(49) *Если Петя ее обидит, она его простит*.

(50) *Даже если <если даже> Петя ее обидит, она его простит*.

Высказывание с *даже*, как и исходное, описывает не реальное и даже не вымышленное положение дел, а гипотезу, предположение говорящего. Тем самым, в семантическую структуру этого примера входят следующие компоненты, выражаемые союзом «*если* гипотезы»:

(51) 'говорящий не знает, какая ситуация имеет место в описываемый отрезок времени; говорящий считает, что в этот отрезок времени возможна [= может иметь место] ситуация P, возможна [= может иметь место] ситуация не-P; говорящий представляет [≈ строит гипотезу]: имеет место ситуация P; говорящий представляет, что в рамках этой гипотезы имеет место следующее'.

Высказывание с *даже если* отличается от высказывания с *если* не этими компонентами, а содержанием гипотезы. Данный смысл представляется так:

(52) (i) 'ситуация 'Петя ее обидит' могла бы повлиять на положение дел; в результате не имела бы место ситуация 'она его простит';

(ii) ситуация 'Петя ее обидит' не влияет на положение дел; в результате имеет место ситуация 'она его простит';

(iii) другие подобные ситуации не влияют на положение дел; можно было с бóльшим основанием ожидать, что другие ситуации не влияют на положение дел, чем то, что ситуация 'Петя ее обидит' не влияет на положение дел'.

В общем виде, *даже если / если даже* в некаузальном контексте выражает следующее значение:

(IV) *Даже если <если даже> P, Q* =

(IVa) ‘говорящий не знает, какая ситуация имеет место в описываемый отрезок времени’;

(IVb) ‘говорящий считает, что в этот отрезок времени возможна [= может иметь место] ситуация P, возможна [= может иметь место] ситуация не-P’;

(IVc) ‘говорящий представляет [≈ строит гипотезу]: имеет место ситуация P’;

(IVd) ‘говорящий представляет, что в рамках этой гипотезы имеет место следующее:

(IVd-i) ‘ситуация P могла бы влиять на положение дел; в результате не имела бы место ситуация Q’;

(IVd-ii) ‘ситуация P не влияет на положение дел; в результате имеет место ситуация Q';

(IVd-iii) другие подобные ситуации не влияют на положение дел; можно было с бо^льшим основанием ожидать, что другие ситуации не влияют на положение дел, чем то, что ситуация P не влияет на положение дел’.

Ясно, что компонент (IVd-iii) — это та модальная семантика, которую выражает частица *даже*. А чему обязаны своим существованием компоненты (IVd-i) и (IVd-ii)? В некаузальной модификации союза «*если* гипотезы» этих компонентов нет. На первый взгляд, перед нами то нестандартное наращение смысла, благодаря которому семантика сочетания *даже если* не сводима к сумме значений его составных частей *даже* и *если*. Однако это не так.

Обсуждаемые компоненты (IVd-i) и (IVd-ii) этого толкования соответствуют довербальному семантическому элементу “псевдозависимость ситуации Q от ситуации P” в разложении «*если* гипотезы». Однако мы видим, что этот компонент существенно изменился и обогатился. Покажем, что это обогащение семантики вполне системно.

В самом общем виде, каузальная модификация указывает на влияние некоторой ситуации на положение дел, а некаузальная модификация такого указания не содержит. Иными словами, довербальный семантический элемент “псевдозависимость ситуации Q от ситуации P” естественно интерпретировать как отсутствие указания на каузальную зависимость.

Противопоставление ‘утверждение A — отсутствие утверждения A’ весьма распространено в системе языка: как заметил Р.Якобсон, оно лежит в основе различения маркированных и немаркированных грамматических категорий [Якобсон 1972]. При этом значение ‘отсутствие утверждения A’ само по себе достаточно бедно — оно не несет никакой позитивной информации. В связи с этим данное значение претерпевает системный сдвиг: ‘отсутствие утверждения A’ — ‘отсутствие A’, т. е. ‘не-A’. Процитируем Р.Якобсона: «Общее значение маркированной категории состоит в утверждении наличия некоторого (положительного или отрицательного) свойства A; общее значение соответствующей немаркированной категории состоит в отсутствии утверждения относительно наличия A; она употребляется главным образом, хотя и не исключительно, для указания на отсутствие A [подчеркнуто мною — Е. У.] <…> Когда речь идет об общем значении данной категории, это противопоставление может быть интерпретировано как «утверждение A — отсутствие утверждения A», тогда как на уровне более узких, специализированных (nuclear) значений мы имеем дело с противопоставлением «утверждение A — утверждение не-A» [Якобсон 1972: 102–103].

Итак, отсутствие указания на каузальную зависимость закономерно интерпретируется как указание на отсутствие такой зависимости. Следовательно, довербальный семантический элемент “псевдозависимость ситуации Q от ситуации P” естественно переосмысляется как указание на отсутствие каузальной зависимости, ср. компонент (IVd-ii) в толковании *если даже*.

Но компонент (IVd-ii) содержит отрицательную пропозицию: ‘ситуация P не влияет на положение дел’. Между тем, хорошо известно, что отрицательное высказывание ‘не-S’, например, *Дождь не кончился*, уместно лишь в случае ожидания, что имеет место S (мы ожидали, что дождь кончился) — иначе нечего отрицать. Иными словами, высказывание ‘не-S’ имеет пресуппозицию: ‘можно было ожидать, что имеет место S’. Значит, такую пресуппозицию имеет и обсуждаемый компонент (IVd-ii) ‘ситуация P не влияет на положение дел’. Эта пресуппозиция наиболее естественно формулируется так: ‘ситуация P могла бы повлиять на положение дел; в результате не имела бы место ситуация Q’. А это и есть компонент (IVd-i) толкования *даже если*.

Итак, хотя в рассмотренных нами некаузальных контекстах семантика *даже если* существенно богаче по сравнению с семантикой составных частей ‘даже’ и ‘если’, однако это обогащение семантики вполне системно и выводимо по общим закономерностям из семантики некаузальной модификации *если*.

Действительно, в некаузальном контексте *Если P, Q* речь идет о том, что ситуация P не влияет на положение дел. Естественным образом, в соответствующем контексте с *даже если / если даже* утверждается, что не только P, но и никакая другая ситуация не влияет на положение дел, — таков вклад в рассматриваемую единицу частицы *даже*.

Перейдем теперь к каузальным контекстам с *если*. Ср. *Если будут пробки, мы опоздаем на самолет*. С логической точки зрения можно было бы ожидать, что в подобных контекстах с *даже если / если даже* речь идет о том, что на положение дел влияет не только данная ситуация P, но и вообще любая ситуация. Иными словами, можно было бы ожидать, что двум модификациям *если* — некаузальной и каузальной — соответствуют и два разных понимания *даже если / если даже*.

Однако язык устроен проще. *Даже если / если даже* во всех контекстах понимается одинаково и указывает на то, что ни P и никакая другая ситуация не влияет на положение дел. Это значит, что в контексте частицы *даже* всегда реализуется некаузальная модификация *если*. Эта некаузальная модификация и в данном случае обогащается по предложенной выше схеме.

Теперь ясно, почему *даже если / если даже* в каузальном контексте отрицательно поляризовано, ср. *Даже если будут пробки, мы не опоздаем на самолет*. Ведь в таком контексте должно быть выражено отрицание пропозиции ‘ситуация влияет на положение дел’. Естественное обозначение отрицания — частица *не*.

Сочетание *даже если* обладает еще одной особенностью: оно, в отличие от союза *если*, тяготеет к контекстам, выражающим непосредственную каузацию, и затруднено в контекстах, обозначающих каузацию косвенную. Отсюда, между прочим, следует необходимость различать выделенные выше типы связей между ситуациями (i)–(iv). Тем самым, анализ *даже если /если даже* верифицирует предлагаемую декомпозицию семантического примитива *если*.

4. Смежные случаи

Внимательное рассмотрение показывает, что частица *даже* может быть введена далеко не в любое даже простое предложение. Ср.

(59) а. *Этот камень поднимет самый сильный человек в мире* VS
*б. *Этот камень поднимет даже самый сильный человек в мире* VS
в. *Этот камень не поднимет даже самый сильный человек в мире*.

(60) а. *Эту задачу слабый ученик не решит* VS
б. **Эту задачу даже слабый ученик не решит* VS
в. *Эту задачу даже слабый ученик решит*.

Дело в том, что частица *даже* ранжирует ситуации на некоторой «шкале ожидания», причем эта шкала согласуется с нашими знаниями о мире. Так, естественно, ожидать, что слабый человек не поднимет камень, и менее естественно ожидать, что его не поднимет самый сильный человек. Нормальный пример (59в) полностью согласуется с этим ожиданием. В аномальном примере (59б) шкала ситуаций, задаваемая *даже*, «перевернута» и поэтому вступает в противоречие с общей энциклопедией. Из-за этого противоречия данный пример абсурден. Точно так же обстоит дело с примерами (60).

В примерах типа (59)–(60) никто не усматривает отрицательно или положительно поляризованных лексем — такие фразы абсурдны потому, что описывают положение дел, которое в принципе не может иметь места и которое поэтому невозможно осмыслить.

Аналогичным образом обстоит дело в сочетаниях *даже когда, даже при условии* и т. п., значение которых равно сумме значений их компонентов.

Иначе обстоит дело с *даже если*. Союз *если* имеет две модификации — каузальную и некаузальную, а сочетание *даже если* всегда понимается некаузально. Данная особенность *даже если* объясняется без привлечения информации о мире: просто в данном сочетании реализуется некаузальная модификация союза *если*.

5. Заключение

Семантика сочетания *даже если / если даже* вполне выводима из значений его составных частей ‘если’ и ‘даже’. Специфика сочетания состоит, прежде всего, в том, что союз *если* в его составе всегда выступает в своей некаузальной модификации. Этот факт должен быть отражен в описании союза *если*. Кроме того, семантика *если* в этом сочетании претерпевает системное обогащение вида: «отсутствие утверждение А» — «утверждение отсутствия А». Тем самым, сочетание *даже если / если даже* может быть описано в словаре в правилах выбора реализации того или иного дизъюнктивного компонента значения союза *если*.

С практической точки зрения *даже если / если даже* часто рассматривают как единицу, входящую в поле языковых средств, выражающих уступительную или условно-уступительную семантику [Грамматика-80; Храковский 2004; В. Апресян 2006]. Подобный функциональный подход не отменяет, однако, теоретических проблем отграничения слова от словосочетания.

Литература

1. *Абаев В. И.* 1934/1995. Язык как идеология и язык как техника // Язык и мышление. Л., 1934. 2. [перепечатано: В. И. Абаев. Избранные труды. Т. II. Общее и сравнительное языкознание. Владикавказ, 1995.]
2. *Апресян Ю. Д.* 1974. Лексическая семантика. М.
3. *Апресян В. Ю.* 2006. Уступительность в языке // Языковая картина мира и системная лексикография / Отв. ред. Ю. Д. Апресян. М.
4. *Богуславский И. М.* 1996. Сфера действия лексических единиц. М.
5. *Вежбицкая А.* 1996а. Язык. Культура. Познание. М.
6. *Вежбицкая А.* 1996б. Семантика «логических понятий» // Московский лингвистический журнал. Т. 2. М..
7. *Грамматика-80* — Русская грамматика. Т. I–II. М., 1980.
8. *Евтюхин В. Б.* 1996. Группировка полей обусловленности: причина, условие, цель, следствие, уступка // Теория функциональной грамматики: Локативность. Бытийность. Посессивность. Обусловленность. СПб.
9. *Жолковский А. К.* 1964. Лексика целесообразной деятельности // Машинный перевод и прикладная лингвистика. Вып. 8 / I Московский государственный педагогический институт иностранных языков. Труды института. М.
10. *Иорданская Л. Н., Мельчук И. А.* 2007. Смысл и сочетаемость в словаре. М.
11. *Крейдлин Г. Е.* 1975. Лексема даже // Семиотика и информатика. М. Вып. 6.
12. *Ляпон М. В.* 1986. Смысловая структура сложного предложения и текст. К типологии внутритекстовых отношений. М.
13. *Мельчук И. А.* 1974. Опыт теории лингвистических моделей «Смысл ⇔ Текст». М.
14. *Падучева Е. В.* 1985. Высказывание и его соотнесенность с действительностью (референциальные аспекты семантики местоимений). М.
15. *Падучева Е. В.* 2004. Динамические модели в семантике лексики. М.
16. *Санников В. З.* 2001/2008. Семантика и прагматика союза если // Русский язык в научном освещении. 2001. № 2. [Воспроизведено в книге: В. З. Санников. Русский синтаксис в семантико-прагматическом пространстве. М., 2008. Часть четвертая, Глава 1 «Конструкции с союзом если». Ссылки на цитаты даны по книге].
17. *Томмола Х.* 2004. Даже если в масле поджарить… (Об условной уступке) // Типологические обоснования в грамматитке. М.
18. *Урысон Е. В.* 2001. Союз если и семантические примитивы // ВЯ. 2001. № 4.
19. *Урысон Е. В.* 2009. Составные союзы А ТО и А НЕ ТО: возможности семантического композиционального анализа// ВЯ, 2009, № 4.
20. *Урысон Е. В.* 2010. Союзы если, когда и раз: попытка сопоставительного семантического анализа // Русский язык в научном освещении. 2010, № 1.
21. *ФЭС* — Философский энциклопедический словарь. М. 1983.
22. *Храковский В. С.* 2004. Типология уступительных конструкций. СПб. 2004.
23. *Якобсон Р.* 1972. Шифтеры, глагольные категории и русский глагол // Принципы типологического анализа языков различного строя. М.
24. *Comrie B.* 1986 Conditionals: a typology // On conditionals / Traugott E., ter Meulen A., Reilly J. Sn., Ferguson A. (eds.). Cambridge.
25. *Ducrot O.* 1972. Dire et ne pas dire. Paris.
26. *Wierzbicka A.* 1997. Conditionals and counterfactuals: conceptual primitives and linguistic universals // On conditionals again / Athanasiadou A., Dirven R. (eds). Amsterdam, Philadelphia.

Экспериментальный подход к исследованию референции в дискурсе: интерпретация анафорического местоимения в зависимости от риторического расстояния до его антецедента[1]

Experimental approach to reference in discourse: effects of rhetorical structure on pronoun intepretation

Федорова О. В. (olga.fedorova@msu.ru),
Деликишкина Е. А. (skaista_diena@mail.ru),
Малютина С. А. (s.malyutina@gmail.com),
Успенская А. М. (ania.quies@gmail.com), **Фейн А. А.** (feinastia@yandex.ru)
Московский государственный университет имени М. В. Ломоносова

1. Введение. Референция в дискурсе: основные понятия

Дискурсивная референция в современной лингвистике является типичным примером относительно автономной области исследований, где представлена обширная и очень разнообразная палитра теоретических моделей и исследовательских подходов. В настоящей работе мы остановимся на одном из таких подходов, связанном с выявлением и описанием различных факторов, потенциально влияющих на выбор говорящим того или иного референциального средства (как то: полная ИГ, анафорическое местоимение или анафорический нуль) и на последующую интерпретацию этой ИГ адресатом сообщения.

В качестве таких значимых факторов в разных исследованиях предлагались как различным образом измеренные расстояния до антецедента — линейное [Givon 1983], риторическое (=РитР, [Fox 1987]) или расстояние в абзацах [Tomlin 1987], так и семантико-синтаксический статус антецедента (подлежащность и агентивность) и его внутренние свойства (одушевленность и центральность в дискурсе). В работах А. А. Кибрика (в частности, [Кибрик 2003], [Kibrik in press]) был предложен многофакторный количественный подход, описывающий процесс референциального выбора, т.е. процесс, ориентированный в первую очередь на говорящего. Согласно этому подходу предполагается, что референциальный выбор зависит от степени активации референта в рабочей памяти говорящего (аналогичные идеи можно найти также в работах У.Чейфа, например, [Chafe 1994]). Таким образом, факторы, оказывающие влияние на референциальный выбор, рассматриваются как факторы активации, которые в сумме дают коэффициент активации (=КА). Чем выше КА, тем более активирован референт и тем больше вероятность употребления в текущем дискурсе редуцированного референциального средства — анафорического местоимения или нуля.

Каждый релевантный фактор активации имеет свой численный показатель, что дает возможность вычислять текущий КА (который обычно находится в пределах от 0 до 1) для каждого референта в каждый момент развертывания дискурса. А. А. Кибрик выделяет несколько значимых порогов активации: например, при КА ниже 0,4 невозможно употребление редуцированной ИГ, а выше порога 0,9, наоборот, — полной ИГ.

Кроме определения степени активации референта в рабочей памяти говорящего описываемая модель референциального выбора включает в себя, в частности, фильтр референциального конфликта (=референциальной неоднозначности), действие которого блокирует использование редуцированного средства в случае высокой активации более чем одного референта.

Данная модель описывает референциальный выбор говорящего. А как происходит интерпретация выбранного говорящим референциального средства в голове у адресата? В настоящем исследовании мы исходим из допущения, что в идеале в референциальном компоненте модели рабочей памяти адресата существует точно такая же референциальная ситуация (которая формируется на основании тех же факторов активации, что и в модели говорящего), и, например, увидев или услышав местоимение *он*, адресат соотносит его с наиболее активированным в своей памяти референтом.

[1] Работа выполнена при частичной финансовой поддержке РГНФ (проект № 08-04-00165а).

2. Исследования дискурсивной референции в экспериментальной лингвистике

Данная работа выполнена в рамках (психо) лингвистического экспериментального подхода, который, несмотря на характерную для него низкую экологическую валидность (результаты, получаемые в лабораторных условиях зачастую плохо экстраполируются на поведение людей в реальной ситуации), дает возможность более четко контролировать экспериментальные условия, формулировать конкретные гипотезы и получать более однозначно интерпретируемые результаты. Хотя исследования, посвященные экспериментальному изучению референции в процессе порождения речи, т. е. с точки зрения говорящего, время от времени тоже появляются (см., например, [Arnold & Griffin 2007]), большинство подобных работ направлено на изучение понимания, т. е. интерпретации референциального выражения адресатом.

Авторы классической работы [Daneman & Carpenter 1980], в которой была введена новая методика определения объема рабочей памяти, которая с тех пор является общепринятой, предположили, что существует корреляция между объемом рабочей памяти человека и его способностью восстанавливать антецедент анафорического местоимения. После прохождения теста на определение объема рабочей памяти испытуемым предлагали прочитать 12 текстовых фрагментов и затем ответить на 2 вопроса, первый из которых относился к некоторому факту из текста, а второй был референциальным. Успешность прохождения теста определялась количеством ошибок в ответах на вопросы. Тексты были созданы так, чтобы расстояние между местоимением и его антецедентом было равно 2, 3, 4, 5, 6 и 7 предложениям. Объем рабочей памяти испытуемых по данной методике принимал значения от 2 до 5. В результате люди с небольшим объемом памяти справились с заданием значимо хуже людей с большим объемом. Так, например, пять человек с объемом памяти 2 правильно ответили в среднем только на 8,2 из 12 вопросов к фактам и только на 5,4 из 12 референциальных вопросов, в то время как шесть человек с объемом рабочей памяти 4 и 5 ответили на 11 из 12 вопросов к фактам и на 9,7 из 12 референциальных вопросов. На основании этих результатов был сделан вывод о том, что существует значимая связь между объемом рабочей памяти и максимальным референциальным расстоянием, при котором испытуемый способен правильно определить референт. Серьезным недостатком данного эксперимента, однако, оказалась неоднородность предложений, которые находились между местоимением и его антецедентом — например, некоторые из них состояли из нескольких клауз, а некоторые были совсем короткими. Также в этом исследовании не учитывались другие факторы, способные потенциально повлиять на успешность определения референта.

В работе [Андреева 2005] линейное расстояние измерялось уже не в предложениях, а в клаузах, а другие факторы активации были закреплены на некотором среднем уровне. Однако полученное в результате небольшое количество референциальных ошибок говорит, скорее всего, о том, что линейное расстояние — недостаточно значимый фактор для того, чтобы самостоятельно влиять на правильность определения референта. С другой стороны, во всех случаях, когда ошибки были совершены, линейное расстояние, как и в [Daneman & Carpenter 1980], превышало объем рабочей памяти испытуемых.

Таким образом, наша задача, в частности, состояла в том, чтобы найти другой, более значимый фактор активации, который и в изоляции смог бы повлиять на правильность определения референта. В качестве такого фактора было выбрано РитР до антецедента редуцированной ИГ.

3. Серия экспериментов на русском материале: фактор риторического расстояния

Итак, в экспериментальном исследовании, описываемом ниже, мы варьировали фактор РитР до антецедента, который, как было показано, в частности, в [Kibrik & Krasavina 2005], является одним из важных факторов активации (все остальные известные факторы активации, которые потенциально могли бы повлиять на правильность определения референта, были закреплены на одинаковом уровне). Традиционно РитР вычисляется по иерархической дискурсивной структуре, которая была разработана в рамках Теории риторической структуры (ТРС, [Mann & Thompson 1988]). Данная теория исходит из постулата о том, что любая единица дискурса связана хотя бы с одной другой единицей данного дискурса посредством некоторой осмысленной связи. В рамках ТРС предлагается метод представления связного текста в виде графа, который позволяет описать структуру текста с помощью ограниченного набора семантических связей (т.н. риторических отношений). Универсальность риторических отношений заключается в том, что они действуют на всех уровнях иерархии и связывают между собой как элементарные дискурсивные единицы (=ЭДЕ, определение и принципы их выделения см., например, в работе [Кибрик & Подлесская 2009: 55 и далее]), так и их группы. Схематично такая иерархическая структура изображается в виде дерева, в узлах которого находятся дискурсивные единицы; наличие риторических отношений между непосредственно связанными единицами обозначается дугами. В данной работе РитР принималось равным количеству го-

ризонтальных переходов по дугам дерева, которые необходимо совершить, чтобы добраться от ЭДЕ, содержащей анафорическое местоимение, до ЭДЕ с его антецедентом (подробнее о данном принципе исчисления РитР см. в работах [Kibrik 1996] и [Kibrik & Krasavina 2005]).

3.1. Создание экспериментальных текстов и листов

Для того, чтобы избежать различий в восприятии разных текстов разными испытуемыми мы искусственно создали шесть текстов длиной чуть менее 100 слов, каждый из которых состоял из общей преамбулы и трех вариантов с РитР в 1, 2, и 3 ЭДЕ, пример такого текста см. в приложении 1; все тексты имели одинаковую риторическую структуру (приложение 2) и одинаковые КА для истинного и двух ложных антецедентов (приложение 3).

Шесть экспериментальных листов были составлены таким образом, что каждый испытуемый читал по два текста с РитР в 1, 2 и 3 ЭДЕ (приложение 4).

3.2. Эксперимент 1: проверка отсутствия референциального конфликта

КА, сосчитанные по методике, предложенной в [Kibrik in press], для двух ложных антецедентов были явно недостаточны для возможности возникновения референциального конфликта. Данный факт был экспериментально подтвержден в ходе нашего первого исследования, в котором 24 испытуемых читали шесть текстов и вопросы к ним, которые показывались на одном листе бумаги. Все испытуемые, отвечая на референциальные вопросы к текстам, однозначно указывали на истинных референтов. Таким образом, можно предположить, что все ошибки, которые возникнут в ходе основного эксперимента, будут связаны с недостаточной активацией референта в рабочей памяти испытуемых.

3.3. Эксперимент 2: проверка естественности текстов

Так как тексты для эксперимента были созданы искусственно, необходимо было проверить их на естественность. Для этого был проведен второй эксперимент, в ходе которого 96 испытуемых оценивали один из трех вариантов каждого текста по пятибалльной шкале (где цифре 0 соответствовало значение «данный текст не является нормальным текстом русского языка», а цифра 4 обозначала «текст взят из хорошей художественной литературы»). Результаты (см. приложение 5) в целом показали некоторый небольшой разброс по текстам, однако все они находились в допустимых пределах по степени сложности восприятия.

Важным результатом эксперимента 2 явился также тот факт, что ни для одного текста не было зафиксировано различий между вариантами с разным РитР, т. е. субъективно увеличение РитР не воспринималось испытуемыми как увеличение сложности.

3.4. Эксперимент 3: роль риторического расстояния

В ходе третьего эксперимента испытуемые читали каждый текст про себя, а затем устно отвечали на появлявшиеся на экране три вопроса, два фактографических и один референциальный[2]. Для того чтобы испытуемый не мог вернуться к уже прочитанной части текста, в которой упоминались ложные и истинный референты, была использована методика саморегуляции скорости, т.е. после нажатия клавиши «пробел» первая часть текста исчезала и вместо нее появлялась вторая часть.

Количество правильных ответов составило 76 % для фактографических вопросов (см. приложение 6) и 52 % для референциальных (приложение 7). По сравнению с результатами из работы [Daneman & Carpenter 1980], где было в среднем 9,4 из 12 (78 %) и 7,4 из 12 (62 %), соответственно, это несколько меньше, однако и средний объем рабочей памяти в том эксперименте был несколько выше (в среднем 3,15 по сравнению с 3 в нашем; на материале больших выборок, средний объем памяти студентов MIT равен 3,25, а по результатам русскоязычного варианта студентов МГУ — 3,2). Необходимо отметить также, что если для испытуемых с небольшим и средним объемами памяти граница между хорошо и плохо понимаемыми текстами проходила между текстами с РитР 1 и 2, то для испытуемых с большим объемом памяти эта граница смещалась вправо (т. е. становилась между РитР 2 и 3). Данные результаты можно трактовать таким образом, что тексты с РитР 1 и 2 являются естественными для восприятия испытуемых с большим объемом памяти, но тексты с РитР 2 уже сложноваты для тех, у кого она меньше; тексты с РитР 3 сложны уже и для испытуемых с большой памятью[3]. Данные результаты хорошо согласуются с подобранными эвристическим путем КА из [Kibrik 1996] и [Kibrik in press] (см. приложение 3).

Еще одним результатом эксперимента явилась выявленная зависимость между РитР и правильностью ответов на референциальные вопросы. Независимо от объема рабочей памяти увеличение РитР всегда оказывало влияние на правильность восста-

---

[2] Сразу после этого эксперимента все испытуемые также проходили тест по определению объема их рабочей памяти.

[3] Более подробно о результатах, связанных с рабочей памятью испытуемых, см. в нашей работе [Федорова и др. 2010].

новления антецедента местоимения (см. приложение 7). Учитывая тот факт, что остальные факторы активации мы закрепили на одном уровне, можно заключить, что этот фактор является принципиально значимым фактором как референциального выбора говорящего, так и интерпретации референциального средства адресатом.

Опишем теперь чуть более подробно, как, по нашему мнению, происходит понимание дискурсивных фрагментов с местоимением *он*. Предположим, на вход рабочей памяти подается фрагмент с *он*, после чего адресат начинает искать в своей текущей референциальной модели наиболее активированный референт; затем результат интегрируется выше, в некоторую ситуационную сеть. Если по каким-то причинам происходит сбой (например, КА оказывается слишком мал), то в этой сети остается лакуна или референт определяется неправильно. В этом случае и возникают тексты-головоломки, когда при ответе на вопрос к антецеденту местоимения человеку приходится выбирать наиболее логичную возможность или просто угадывать, но никак не восстанавливать его из памяти. Таким образом, дискурсивная сложность данного типа сильно отличается от других видов языковой сложности (например, известных предложений с object- *vs*. subject-relativization), здесь вообще нет сложности как таковой (если измерять сложность в количестве вычислительных ресурсов, необходимых для обработки сообщения) — при ответе на референциальный вопрос испытуемый просто не способен правильно восстановить референт, если он был недостаточно активирован, так как тот фрагмент, в котором было его предыдущее упоминание, уже давно был обработан.

Наконец, важным отличием наших данных от результатов предыдущих экспериментов оказался тот факт, что при увеличении РитР количество ошибок возрастает только для референциальных вопросов, но не фактографических (ср. приложение 6 и 7). Данные результаты логично трактовать в пользу существования отдельного, относительно автономного референциального модуля, отвечающего за порождение референциальных выражений говорящим и за понимание их адресатом.

4. Заключение

Итак, по результатам трех проведенных экспериментов можно сделать следующие основные выводы: (i) за процесс референции отвечает особый референциальный модуль; (ii) успешность интерпретации референциального выражения зависит от объема рабочей памяти испытуемых; (iii) субъективно увеличение РитР не воспринимается испытуемыми как увеличение сложности; (iv) РитР является значимым фактором активации, который при этом оказывается достаточно сильным, чтобы самостоятельно влиять на понимание референциальных выражений; (v) КА, подобранные эвристическим путем, хорошо коррелируют со способностью адресатов (с разными объемами рабочей памяти) восстанавливать антецеденты анафорического местоимения в экспериментальном исследовании.

Однако уже после проведения третьего эксперимента стало понятно, что наши тексты были не совсем естественны с точки зрения их деления на абзацы (см. приложение 1), что теоретически могло повлиять на интерпретацию их испытуемыми (скорее всего, в сторону увеличения количества ошибок). Таким образом, наша следующая задача состоит в подтверждении полученных результатов на более экологически валидном экспериментальном материале.

Приложение 1. Пример стимульного материала

Был конец рабочего дня. Пятая бригада скорой помощи ехала на базу после ложного вызова. На носилках в кабине, набегавшись за смену, прикорнул **медбрат**. Усталый **доктор**, слушавший музыку в плеере, игнорировал заискивающие взгляды молодого **ассистента**, горящего рабочим энтузиазмом после первого дня в бригаде.

В наушниках звучал «Белый альбом» битлов. Безупречная мелодия качала и убаюкивала.

а) РитР=1 **Он** испытывал легкие угрызения совести за свою невнимательность к коллеге, но усталость превозмогала всё.

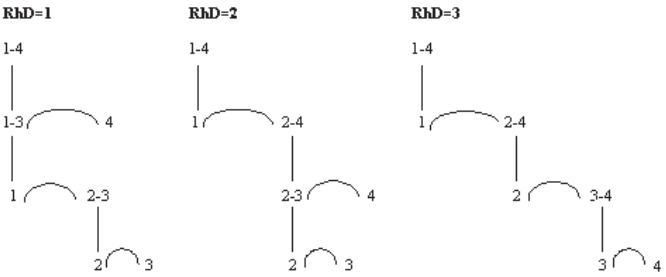
б) РитР=2 **Он** любил слушать эту пластинку после тяжелого трудового дня.

в) РитР=3 **Он** почувствовал, что медленно проваливается в сон.

Вопросы:

1. Какой номер был у бригады скорой?
2. Какая запись звучала в плеере?
3. а) Кому было стыдно за невнимательность к коллеге? б) Кто любил слушать пластинку после тяжелого трудового дня? в) Кто почувствовал, что засыпает?

Приложение 2. Деревья ТРС к текстам



Приложение 3. **Коэффициенты активации** (подсчитаны по работе [Kibrik in press])

КА без учета РитР:

- **ложного антецедента, встречающегося в тексте первым**:
  AS − RhD (медбрат) = −0.3 (LinD = 4) + 0.2 (animacy: Human) − 0.1 (introductory referent: Yes) − 0.3 (ParaD=1) = ***−0.5***
- **ложного антецедента, идущего в тексте последним**:
  *AS − RhD (ассистент) = −0.2 (LinD = 3) + 0.2 или 0 (synt role: DO или Other) + 0.2 (animacy: Human) − 0.3 (ParaD = 1) = −****0.1*** *или* ***−0.3***
- **истинного антецедента**:
  AS − RhD (доктор) = −0.2 (LinD = 3) + 0.4 (synt role: S of a main clause) + 0.2 (animacy: Human) − 0.1 (introductory referent: Yes) − 0.3 (ParaD = 1) = ***0***

Полный КА истинного антецедента (*доктор*):
AS (RhD 1) = 0 + 0.7 = **0.7** **полная ИГ или местоимение**
AS (RhD 2) = 0 + 0.5 = **0.5** **более вероятно полная ИГ, но возможно и местоимение**
AS (RhD 3) = 0 + 0 = **0** **только полная ИГ**

Приложение 4. Распределение текстов по экспериментальным листам

(тексты обозначаются маленькими латинскими буквами, РитР — цифрами 1, 2 и 3)
лист 1: a1, d2, c2, f3, e3, b1
лист 2: b1, e3, f3, c2, d2, a1
лист 3: c3, f1, d3, a2, b2, e1
лист 4: e1, b2, a2, d3, f1, c3
лист 5: f2, a3, e2, b3, c1, d1
лист 6: d1, c1, b3, e2, a3, f2

Приложение 5. Результаты эксперимента 2

текст / РитР	1	2	3	среднее
a	2,8	2,4	2,6	2,6
b	2,5	2,4	2,5	2,4
c	3,2	3,2	2,9	3,1
d	2,3	2,5	2,3	2,4
e	2,5	2,6	2,6	2,5
f	2,9	2,5	2,9	2,8
среднее	2,7	2,6	2,7	2,7

Приложение 6. Результаты эксперимента 3: правильность ответов на фактографические вопросы (в %)

кол-во исп.	объем памяти / РитР	1	2	3	среднее
86	2–3 (мал.)	74	70	74	73
17	3,5–4 (средн.)	78	74	75	76
17	4,5–5 (больш.)	94	94	89	92
120	все вместе	78	74	76	76

Приложение 7. Результаты эксперимента 3: правильность ответов на референциальные вопросы (в %)

кол-во исп.	объем памяти / РитР	1	2	3	среднее
86	2–3 (мал.)	62	38	28	43
17	3,5–4 (средн.)	85	68	41	65
17	4,5–5 (больш.)	100	94	68	73
120	все вместе	70	50	35	52

Литература

1. *Андреева К. В.* Влияние объема оперативной памяти человека на его языковое поведение (на примере определения референта анафорического местоимения). Неопубликованная дипломная работа. МГУ, 2005.
2. *Кибрик А. А.* Анализ дискурса в когнитивной перспективе. Дисс. доктора филол. наук. М.: Институт языкознания РАН, 2003.
3. *А. А. Кибрик, В. И. Подлесская (ред.)* Рассказы о сновидениях: Корпусное исследование устного русского дискурса. М.: ЯСК, 2009.
4. *Федорова О. В., Деликишкина Е. А., Малютина С. А., Успенская А. М., Фейн А. А.* Совместное или раздельное использование вербальных ресурсов рабочей памяти в процессе понимания предложений // Четвёртая международная конференция по когнитивной науке, Томск, 2010.
5. *Arnold J. E., Griffin Z.* The Effect of Additional Characters on Choice of Referring Expression: Everyone Competes // Journal of Memory and Language, 2007, 56. P. 521–536.
6. *Chafe W.* Discourse, consciousness, and time. The flow and displacement of conscious experience in speaking and writing. Chicago: University of Chicago Press, 1994.
7. *Daneman M., Carpenter P.* Individual differences in working memory and reading // Journal of Verbal Learning and Verbal Behavior, 1980, 19. P. 450–466.
8. *Fox B.* Discourse structure and anaphora in written and conversational English. Cambridge: Cambridge University Press, 1987.
9. *Givón T. Topic continuity in spoken English // T. Givón* (ed.) Topic continuity in discourse: A quantitative cross-language study. Amsterdam: Benjamins, 1983. P. 345–363.
10. *Kibrik A. A., Krasavina O. N.* A corpus study of referential choice: The role of rhetorical structure // И. М. Кобозева, А. С. Нариньяни, В. П. Селегей (ред.) Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции Диалог 2005. М.: Наука, 2005. С. 561–569.
11. *Kibrik A. A.* Anaphora in Russian narrative discourse: A cognitive calculative account // B. Fox (ed.) Studies in anaphora. Amsterdam: Benjamins, 1996. P. 255–304.
12. *Kibrik A. A.* Reference in grammars, discourse, and cognition. Oxford: Oxford University Press, in press.
13. *Mann W. C., Thompson S. A.* Rhetorical structure theory: toward a functional theory of text organization // Text, 8, 1988. P. 243–281.
14. *Tomlin R. S.* Linguistic reflections of cognitive events // R. Tomlin (ed.) Coherence and grounding in discourse. Amsterdam: Benjamins, 1987. P. 455–479.

Автоматическая расстановка пауз в системе синтеза русской речи по тексту

Automatic pause placement in a Russian text-to-speech system

Хомицевич О. Г. (khomitsevich@speechpro.com),
Соломенник М. В. (solomennik-m@speechpro.com)

ООО «Центр речевых технологий», Санкт-Петербург

В статье описывается алгоритм расстановки пауз в синтезированной речи (система синтеза речи VitalVoice), основанный на определении синтаксических связей слов в предложении. Рассматриваются результаты работы алгоритма на текстах разных типов и правильность паузации по сравнению с базовой системой синтеза речи, разработанной ранее.

1. Введение

Одной из важных задач при создании автоматического синтеза речи по тексту является расстановка пауз в синтезируемом тексте. Корректная паузация необходима для комфортности восприятия речи, а во многих случаях — и для правильного понимания смысла предложения. Значительная часть информации о паузах передается в тексте с помощью знаков препинания, однако могут встретиться и большие отрезки без всяких знаков, которые с трудом воспринимаются на слух при прочтении без пауз. С другой стороны, наличие знака препинания внутри предложения не всегда говорит о необходимости или возможности паузы. Таким образом, полноценная система синтеза речи по тексту нуждается в специальном алгоритме деления на синтагмы и расстановки пауз, который бы учитывал не только знаки препинания, но и лексический и синтаксический контекст. В статье будет рассмотрен алгоритм нахождения мест пауз в синтезируемом тексте, реализованный в системе синтеза русской речи «Vital Voice» [9] (разработка компании «Центр речевых технологий»).[1]

[1] Нахождение синтагматических границ в тексте не обязательно означает расстановку пауз на всех границах синтагм; однако, поскольку в нашей системе синтеза речи все интонационные границы снабжены паузами той или иной длины, в дальнейшем мы будем говорить о паузации и о синтагматическом членении, имея в виду один и тот же механизм.

2. Способы расстановки пауз в синтезированном тексте

Деление письменного текста на синтагмы может выполняться различными способами: применяются как статистические, так и детерминистские (основанные на наборе заранее заданных правил) подходы к данной задаче. Статистические способы выделения синтагм имеют достаточно высокую эффективность [2]; однако трудность применения их к русскому материалу заключается в том, что для тренировки статистической модели необходим достаточно большой текстовый корпус, размеченный интонационно и со снятой омонимией (в том случае, если при подсчете статистики применяются грамматические, например частеречные, характеристики слов). В то же время, механизмы, основанные на заранее заданных экспертами правилах, могут быть излишне жесткими и не поддающимися изменению и адаптации. Алгоритм паузации, использующийся в системе синтеза русской речи по тексту «Vital Voice», является детерминистским, то есть основан на правилах, разработанных на основании исследований закономерностей синтагматического членения в русском языке. Однако гибкость алгоритма обеспечивается доступностью основных правил паузации для редактирования разработчиком-лингвистом: правила могут быть с легкостью изменены при необходимости исправления обнаруженных ошибок или учета новых, ранее не рассматривавшихся случаев. Данная система

является развитием алгоритмов расстановки пауз, содержавшихся в системе синтеза речи по тексту «Оратор» (разработка компании «Центр речевых технологий», 2005 г.) [6]. В ходе разработки новой системы анализа текста для синтеза речи исходный механизм расстановки пауз был существенно доработан, что привело к значительному повышению правильности расстановки пауз.

3. Основы рассматриваемой системы расстановки пауз

Система расстановки пауз, изначально разработанная в рамках системы синтеза речи по тексту «Оратор», включает в себя несколько этапов анализа текста для определения мест синтагматических границ. В первую очередь учитываются знаки препинания: в большинстве случаев, пунктуационные знаки соответствуют границам синтагм. При этом в исходной системе паузации «Оратора» имелось небольшое количество правил для тех случаев, когда, несмотря на наличие знаков пунктуации, пауза в определенном месте текста нежелательна или невозможна.

На следующем этапе происходит поиск возможных мест пауз в последовательностях слов, не разделенных знаками препинания. При этом анализируются только те отрезки, которые оказались длиннее критического значения (для системы паузации «Оратора» это пять слов). Кроме того, рассматривается небольшой список «неделимых» последовательностей слов (например, «так как»). При нахождении последовательности слов из этого списка, постановка паузы между ними невозможна.

Основой поиска возможных мест пауз между словами является исследование синтаксических связей слов. В то же время, работа системы основана на предположении, что для правильного деления предложения на синтагмы не нужно производить полный анализ синтаксической структуры: достаточно выделить самостоятельные группы слов, между которыми в принципе возможна постановка паузы, а внутри которых пауза маловероятна. Поиск таких групп слов осуществляется при помощи сопоставления словам синтаксических шаблонов — заранее заданных последовательностей частей речи и/или грамматических форм, соответствующих различным часто встречающимся в текстах словосочетаниям. При построении системы шаблонов учитываются следующие частеречные категории: личная форма глагола, инфинитив, существительное, полное и краткое прилагательное, наречие, предлог, (количественное) числительное, личное местоимение. Также учитываются грамматические характеристики слов: род, число, падеж и др., а также согласование между различными частями речи. Дополнительные характеристики включают отдельные семантические признаки слов, а также возможность задания правила для конкретного слова. Набор шаблонов содержится в отдельных файлах, доступных для редактирования лингвистом. Ниже приводится пример шаблона:

$$V + I + PREP(accord=4) + N$$

Данное правило состоит из четырех элементов и применяется к последовательности, начинающейся с личной формы глагола, за которой следует инфинитив, далее предлог, согласованный с четвертым словом шаблона (существительным), и существительное. Этот шаблон будет найден, к примеру, в предложении «Представитель администрации президента предпочел воздержаться от комментариев на эту тему»: группа слов «предпочел воздержаться от комментариев» будет выявлена как объединенная синтаксическими связями.

Для каждого рассматриваемого слова ищутся все последовательности, которые начинаются с него в данном предложении. В итоге для всего рассматриваемого отрезка предложения (от одного знака препинания до другого) мы имеем набор частеречных групп, причем эти группы могут быть вложенными или пересекающимися. Очевидно, что нахождение максимального количества синтаксических связей в предложении еще не означает, что автоматически будет получена информация о потенциальных разрывах синтагм. Напротив, наиболее сложным элементом алгоритма является интерпретация полученного набора связей и поиск такой позиции, в которой разрыв синтагмы является наиболее вероятным. Таким образом, для правильности работы алгоритма важен как общий набор шаблонов, заданный в программе, так и механизм интерпретации полученного для конкретного предложения множества шаблонов.

В системе нахождения мест пауз, использовавшейся в программе синтеза речи «Оратор», производилась оценка возможности паузы внутри каждой пары слов в рассматриваемом отрезке предложения. В алгоритме принимались во внимание два основных фактора — длина группы и вес связи, который в свою очередь зависел от количества групп, в которые входят оба рассматриваемых слова, и количества групп, с которых начинается или на которые заканчивается каждое из слов. Более длинным группам присваивался больший вес, при этом существовали пороговые значения количества пересекающихся групп и их длины, при достижении которых разрыв становился невозможен.

Результаты работы алгоритма, приведенные в публикациях [4, 5, 8], демонстрируют достаточно высокий процент правильно определенных мест разрывов. Однако при разработке новой системы синтеза речи с высокими требованиями к есте-

ственности звучания возникла необходимость исправить имевшиеся у системы паузации недостатки и усовершенствовать ее. Так, распространенной проблемой системы паузации «Оратора» было недостаточное количество пауз в длинном отрезке текста по причине нахождения слишком большого количества длинных пересекающихся групп слов (на основе частеречных шаблонов). Однако простое ослабление условий постановки разрыва приводило к появлению множества ошибок на других предложениях. В связи с этим было решено изменить логику работы с шаблонами; новые принципы работы алгоритма изложены в следующей секции.

4. Новый алгоритм расстановки пауз

Усовершенствование алгоритма расстановки пауз в системе синтеза речи производилось по нескольким направлениям. Во-первых, логика деления отрезка предложения на паузы была изменена и сделана более гибкой. Во-вторых, были произведены количественные улучшения в алгоритме: дополнен набор элементов, рассматривающихся особым образом, а также сам набор частеречных шаблонов.

4.1. Учет веса групп

В текущем алгоритме паузации было решено отказаться от учета длины группы при определении возможного места разрыва. Это связано с тем, что длина шаблона, заданного в правилах, не всегда коррелирует с реальным количеством связанных между собой слов — их может быть значительно больше, и тогда группа вполне может распасться на две или более синтагмы. К тому же, степень близости синтаксической связи в той или иной группе зависит от ее состава, а не от длины. В этом смысле, алгоритм был упрощен: при определении места паузы учитывается только количество групп, в которые входят два соседних слова, но не длина этих групп.

Параметризация весов групп была вынесена из программной части алгоритма в сами файлы правил: внутри групп было выделено два типа синтаксической связи. Связь первого типа может быть разорвана, если соответствующая позиция между двумя словами будет определена как возможное место паузы; связь второго типа не может быть разорвана ни при каких обстоятельствах. К связям второго типа относятся, например, отношения между предлогом и существительным, согласованным определением и существительным, и т. п. Таким образом, при поиске возможного места паузы внутри группы слов, соответствующей частеречному шаблону, учитываются только позиции возможных разрывов, а позиции, соответствующие близкой синтаксической связи, игнорируются. Это позволяет избежать наиболее грубых ошибок паузации, когда паузой разрываются тесно связанные друг с другом члены предложения.

4.2. Изменение состава частеречных шаблонов

Отладке подвергся также состав частеречных шаблонов, перенесенный в новую систему синтеза речи из старой. Изначально набор шаблонов был определен в ходе анализа большого объема текстов, в результате чего были выделены наиболее часто встречающиеся последовательности грамматических категорий. Однако повторяемость в тексте определенной последовательности членов предложения еще не означает, что такая последовательность не будет разрываться паузами. Так, вполне естественно звучит пауза между группой подлежащего и группой сказуемого, перед некоторыми видами предложных групп, даже несмотря на наличие синтаксической связи между соседними словами (ср.: «депутат областного парламента Сергей Иванов / перешел на работу к новому губернатору»; «дизайнер продемонстрировал солидарность / с британскими и американскими модельерами»). С учетом этого, из состава шаблонов были удалены те, которые часто разрываются паузой, либо в них ставилась связь первого типа (с возможностью разрыва).

С другой стороны, набор шаблонов был дополнен новыми видами связанных групп, встречающимися в текстах. Так, необходимым оказалось ввести значительное количество групп, отражающих дистанционные синтаксические отношения между словами в предложении (например, «довольный своей курсовой работой студент»). С помощью системы анализа текста, не предполагающей полного синтаксического разбора предложения, обнаружить синтаксические отношения между словами, не примыкающими друг к другу, крайне сложно, однако некоторые частотные случаи были зафиксированы в виде шаблонов, что позволило выделять их в потоке текста и учитывать при паузации.

4.3. Учет специальных случаев возможности или невозможности постановки пауз

Алгоритм с использованием частеречных шаблонов, несмотря на свою эффективность в учете синтаксических связей между словами, не позволяет учесть многие частные случаи возможности/невозможности синтагматического разрыва в определенных позициях. Отдельные алгоритмы были созданы для учета следующих конструкций и выражений:

- Однородные члены предложения (например, «пять, шесть или семь»);
- Вводные слова («например», «наверно», «в частности» и т. п.);
- Междометия («ну», «ах» и т. п.);
- Обращения (например, «До свидания, господа»);
- Сложные предлоги, энклитики, послелоги, союзы («несмотря на», «спустя», «уж», «бы» и т. п.);
- Обозначения дат, времени (например, «двадцатого мая две тысячи десятого года»);
- Неразрывные идиоматические последовательности («задом наперед», «бог весть» и т. п.)

В некоторых из вышеперечисленных случаев, постановка паузы в месте наличия запятой является неправильной (например, при выделении запятыми слова «пожалуйста» или «увы»). В других случаях, таких как идиоматические выражения, связь между словами не может быть описана с помощью частеречного шаблона и задается специально для конкретных слов.

4.4. Примеры работы алгоритма

Рассмотрим примеры обработки предложения, не содержащего знаков препинания.

(5) *По сравнению с предыдущими семью неделями рост несколько замедлился.*

В данном предложении системой обнаружения синтаксически связанных групп (шаблонов) будут найдены следующие группы: *по сравнению*, *предыдущими семью неделями*, *предыдущими семью*, *семью неделями*, *рост несколько затормозился*, *несколько затормозился*. При этом группы слов *предыдущими семью неделями*, *несколько затормозился* определяются как связанные близкой синтаксической связью (согласованное прилагательное и числительное с существительным; наречие с глаголом). Таким образом, они не могут быть разделены паузой. Кроме того, словосочетание «по сравнению с» будет неразрывно связано с последующим прилагательным, поскольку оно находится в списке сложных предлогов. Единственная пара слов, между которыми не нашлось никакой синтаксической связи, — «неделями рост». Между этими словами будет проставлена пауза. Наглядная схема разбора предложения приведена на рис. 1 (частеречные шаблоны даны в крайнем правом столбце; связь первого типа, которая может быть разорвана, обозначена «+», связь второго типа (неразрывная) — «&»).

(6) *Перезагрузка должна стать длительным усилием со стороны американского и российского народа.*

Для данного предложения обнаруживаются следующие группы слов, соответствующие частеречным шаблонам: *перезагрузка должна стать*, *должна стать*, *стать длительным усилием*, *стать длительным*, *длительным усилием*, *длительным усилием со стороны*, *усилием со стороны*, *со стороны*, *стороны американского и российского народа*, *американского и российского народа*, *американского и российского, российского народа*. Таким образом, между всеми словами предложения существует связь, определенная наличием соответствующего частеречного шаблона, и алгоритм должен решить вопрос, в каком месте связь может быть разорвана.

Из множества пар слов, которые могут быть разделены паузой, исключается *перезагрузка должна*, поскольку единственное слово, не выделенное знаками препинания, в текущем алгоритме не выделяется в отдельную синтагму. Далее, внутри ряда частеречных групп между словами задана неразрывная связь: *должна стать* (краткое прилагательное и инфинитив), *стать длительным* (вспомогательный глагол и прилагательное в твор.пад.), *длительным усилием*, *российского народа* (согласованное прилагательное и существительное), *со стороны* (предлог и существительное). Эти слова не могут быть разделены паузой. Также определяем как неразрывное сочетание *американского и российского* (однородные члены, соединенные союзом *и*). *Со стороны* определяется как принадлежащее к списку сложных предлогов и «связывается» с последующим прилагательным. В то же время внутри групп *усилием со стороны* и *дополнительным усилием со стороны* связь между существительным и предлогом определена как связь «первого» типа, которая может быть разорвана. Таким образом, алгоритм определяет пару слов *усилием со* как возможное место синтагматической границы. Между этими словами и будет поставлена пауза. Схема разбора приведена на рис. 2.

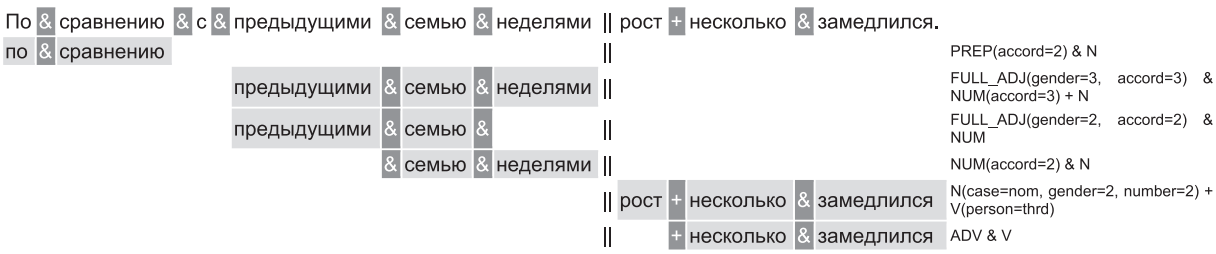


Рис. 1. Схема разбора предложения (1).

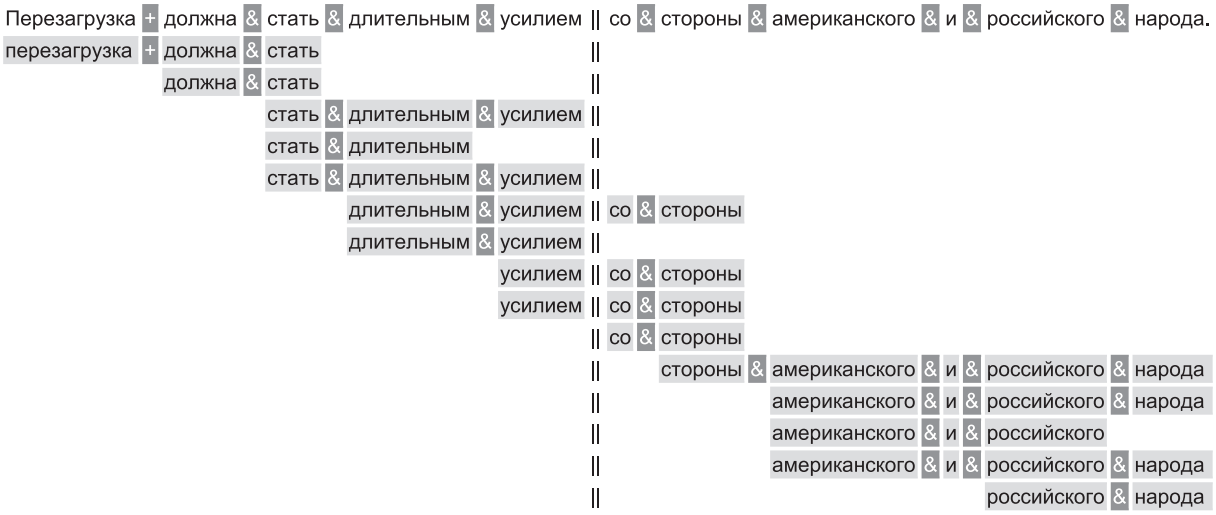


Рис. 2. Схема разбора предложения (2).

5. Результаты работы системы паузации

Оценка результатов работы алгоритма деления на синтагмы в текущей системе синтеза речи показывает улучшение по сравнению с результатами работы базовой системы паузации, реализованной в программе «Оратор». Следует отметить, что оценка правильности расстановки пауз в тексте сама по себе является сложной задачей, в первую очередь потому, что места постановки пауз не определены строго, и деление на синтагмы может сильно варьироваться у разных говорящих даже при чтении одного и того же предложения.

Для оценки адекватности работы алгоритма используются четыре параметра оценки, определенные в работах [1, 2]:

- правильность разрывов (BC=breaks correct) — отношение правильно определенных мест разрывов к общему количеству разрывов в тексте;
- правильность связей (JC=junctures correct) — правильность определения наличия/отсутствия разрыва для каждой пары слов;
- вставка разрыва (JI=juncture insertion) — «лишние» разрывы, вставленные алгоритмом;
- удаление разрыва (JD=juncture deletion) — разрывы, не найденные алгоритмом.

В публикации [5] приведены результаты тестирования системы «Оратор» по корпусу спонтанной речи. При этом сравнивалась правильность постановки пауз системой синтеза и реальные паузы, которые звучали в речи дикторов. В тексте, поданном на вход системы, были проставлены знаки препинания. Это облегчает задачу системы, поскольку она изначально ориентирована на текст, содержащий пунктуацию. В таблице *1* приведены результаты, опубликованные в [5], в сравнении с результатами, полученными на аналогичном тесте для новой системы паузации. Таким образом, количество ошибочно поставленных пауз сократилось почти в два раза, при незначительном росте количества мест, оставленных без пауз.

Таблица 1. Сравнение результатов работы алгоритмов на материале корпуса разговорной речи INTAS [3]

	JC (%)	**BC** (%)	**JI** (%)	**JD** (%)
Система «Оратор»				
Среднее значение	92,05	78,12	4,08	3,87
Станд. откл.	1,49	5,50	1,24	1,10
Система «VitalVoice»				
Среднее значение	93,19	83,48	2,33	4,48
Станд. откл.	1,39	4,63	1,21	1,40

Описанная выше система оценки используется для разных типов алгоритмов (статистических и детерминистских); однако она не позволяет полностью учесть вариативность расстановки пауз, поскольку тексты, применяемые для тестирования, размечаются так, что допускают лишь наличие или отсутствие паузы в конкретном месте.

Наши исследования, проведенные на одних и тех же текстах, прочитанных разными дикторами, показали, что существует достаточно много мест (порядка 19 %) в которых наличие или отсутствие паузы жестко не определено. Это подтверждают и наблюдения других авторов [7]. Чтобы учесть эту вариативность, система оценки была модифицирована; места возможных пауз в текстах, на которых велось тестирование, были помечены тремя различными маркерами: обязательная пауза, обязательное место отсутствия паузы, возможное место паузы

(на основании постановки пауз разными дикторами). Таким образом, если пауза, сгенерированная системой, попадала в место возможной паузы, то эта пауза считалась правильной, также правильной считалась ситуация, если в месте возможной паузы пауза отсутствовала. Так как есть места, в которых наличие паузы факультативно, то и общее количество пауз в тексте определено в пределах от количества обязательных мест пауз до суммы обязательных мест пауз и возможных мест пауз. Поэтому параметр BC не рассчитывается, а вместо него вводятся два дополнительных параметра CP — количество правильно поставленных пауз от общего количества проставленных пауз и CJ — количество правильных мест без пауз от общего количества мест, оставленных без пауз.

В таблице *2* приведены результаты работы систем «Оратор» и «Vital Voice» на одном и том же художественном тексте с учетом необязательных мест пауз. Для анализа работы систем использовался текст Ю.Трифонова «Обмен», начитанный восемью дикторами, что позволило сделать разметку теста для анализа всех возможных и обязательных мест пауз и их отсутствия. Данные таблицы показывают, что количество ошибочно поставленных пауз в новой системе паузации сократилось почти в два раза и почти в три раза сократилось количество мест, ошибочно оставленных без пауз.

Таблица 2. Результаты работы систем «Оратор» и «Vital Voice» с учетом необязательных мест пауз

JC (%)	**CP** (%)	**CJ** (%)	**JI** (%)	**JD** (%)
Система «Оратор»				
97,88	98,84	97,57	1,84	0,28
Система «VitalVoice»				
98,90	99,59	98,68	1,00	0,10

В таблице 3 приведено сравнение работы систем на тексте новостей, с использованием разметки пауз, поставленных двумя разными дикторами. В данном случае при расчете учитывалось только два возможных значения — пауза есть и пауза отсутствует. Сравнение этих данных с таблицей *1* говорит о том, что тексты новостей являются более сложными для правильной расстановки пауз. Возможно, это связано с тем, что в таком тексте содержатся большие отрезки без знаков препинания, а также большое количество слов, не содержащихся в словаре, грамматические характеристики которых неизвестны системе. Тем не менее, для новой системы паузации количество ошибочно поставленных пауз сократилось в полтора раза и в среднем на 15% сократилось количество мест, ошибочно оставленных без пауз.

Таблица 3. Сравнение работы систем для новостей, начитанных разными дикторами

	JC (%)	**BC** (%)	**JI** (%)	**JD** (%)
Система «Оратор»				
Диктор 1	86,58	65,14	5,39	8,03
Диктор 2	87,96	69,60	5,56	6,48
Система «VitalVoice»				
Диктор 1	89,59	70,03	3,47	6,94
Диктор 2	90,59	73,98	3,83	5,57

6. Проблемы и задачи для дальнейшей работы

Применение существующего алгоритма расстановки пауз связано с некоторыми ограничениями, которые могут приводить к ошибкам паузации. Так, отсутствие полного синтаксического анализа предложения обеспечивает более гибкий подход к анализу отдельных частей предложения; как правило, частичного синтаксического анализа бывает достаточно, по крайней мере, для определения мест, где пауза невозможна. Однако некоторые дистанционные синтаксические отношения невозможно определить без более глубокого анализа; также не учитывается взаимное расположение выделенных синтаксических групп, которое в некоторых случаях может влиять на паузацию.

Зачастую ошибки, обнаруживающиеся в работе алгоритма паузации, связаны с распространенной в языке синтаксической неоднозначностью, являющейся проблемой для любого вида синтаксического анализа. Например, неоднозначность присоединения предложных групп может вести к неадекватной паузации перед предлогом (пауза, как правило, нежелательна, если предложная группа относится непосредственно к предшествующему слову, но может быть поставлена, если группа относится к другому слову, отделенному от предлога, ср: «Люксембург поддерживает кандидатуру России /(?) для вступления /(?) во Всемирную торговую организацию»). Такого рода неоднозначность не может быть разрешена в рамках имеющегося алгоритма, так как он определяет синтаксические отношения только в пределах отдельного частеречного шаблона; кроме того, для альтернативных грамматических характеристик могут быть применены разные шаблоны, причем выбор между такими конфликтующими шаблонами не может быть осуществлен из-за отсутствия анализа синтаксической структуры на более высоких уровнях. (Следует заметить, что часть синтаксической неоднозначности снимается в системе синтеза речи

на предыдущем этапе разбора, при снятии омонимии/омографии).

Дальнейшее развитие алгоритма должно включать в себя отладку количества и состава частеречных шаблонов, а также оптимизацию обработки сложных случаев членения предложения на синтагмы при наличии таких элементов, как вводные слова, обращения и др. Для качественного улучшения снятия синтаксической неоднозначности будет необходимо обращение к более глубокому синтаксическому анализу.

7. Выводы и заключение

В статье был описан алгоритм деления письменного текста на синтагмы, использующийся в системе синтеза речи, которая была разработана компанией «Центр речевых технологий». По сравнению с базовой системой синтеза речи, разработанной ранее, алгоритм был усовершенствован и показывает повышение правильности паузации. Использование алгоритма в системе синтеза речи обеспечивает естественность и комфортность восприятия синтезируемой речи. В статье также предложен метод оценки правильности расстановки пауз с учетом необязательных мест пауз.

Литература

1. *Atterer M.* Assigning Prosodic Structure for Speech Synthesis: A Rule-based Approach // Proc. of Prosody 2002, Aix-en-Provence. P. 147–150.
2. *Black A. W., Taylor P.* Assigning phrase breaks from part-of-speech sequences // Computer Speech & Language, Volume 12, Number 2, April 1998. New York: Academic Press — P. 99–117.
3. *Bondarko L. V.* [et al.] Phonetic Properties of Russian Spontaneous Speech // Proceedings of the 15th ICPhS. Barcelona, Spain — 2003. P. 2973–2976.
4. *Oparin I.* Flexible rule-based breaks assignment for Russian // SPECOM 2005 proceedings: 10th international conference Speech and computer: 17–19 October, 2005. Patras, Greece. — Moscow: Moscow State Linguistics University, 2005. — P. 293–296.
5. *Oparin I.* Robust rule-based method for automatic break assignment in Russian texts // Text, speech and dialogue: 8th international conference, TSD 2005: Karlovy Vary, Czech Republic, September 12–15, 2005 : proceedings. — Berlin: Springer, 2005. P. 356–363.
6. *Вольская Н.* и др. Синтезатор русской речи по тексту нового поколения // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог'2005» (Звенигород, 1–6 июня, 2005 г.). Под ред. И. М. Кобозевой, А. С. Нариньяни, В. П. Селегея. М.: Наука, 2005. С. 234–237.
7. *Кривнова О. Ф., Чардин И. С.* Паузирование в естественной и синтезированной речи. — Донецк, 2002. — http://www.russian.slavica.org/article9348.html.
8. *Опарин И. В.* Автоматическое интонационное членение для решения прикладных задач // Интегральное моделирование звуковой формы естественных языков. СПб., 2005. С. 54–64.
9. *Система* синтеза русской речи по тексту VitalVoice: http://vitalvoice.speechpro.com/.

Извлечение информации из текстов на конференциях серии диалог: взгляд соседа по лестничной клетке

Information extraction at dialogue conferences: aview from the neibourhood

Хорошевский В. Ф. (khor@ccas.ru)

Вычислительный центр им. А. А. Дородницына РАН, Москва

В работе представлен ретроспективный анализ публикаций конференций серии Диалог за 2000–2009 г. г. в области извлечения информации из текстов. Для построения общей «картины мира» конференций этой серии используются статистические методы, методы онтологического инжиниринга и семантической кластеризации и классификации, а для выявления значимых семантических пространств — технологии извлечения информации из текстов.

1. Немного истории вместо введения

Как известно, история конференций серии Диалог своими корнями уходит в 80-е годы прошлого столетия, когда в нашей стране под эгидой Научного совета по проблеме «Искусственный интеллект» при Президиуме АН СССР был сформирован междисциплинарный проект «Диалог», в рамках которого стали регулярно проводиться небольшие, но очень значимые по составу участников научные семинары. У истоков организации этих микроконференций стояли такие известные ученые, как Ю. Д. Апресян, А. П. Ершов, А. Е. Кибрик, Г. С. Поспелов и др., а безусловными лидерами, вокруг которых и «раскрутилась» вся эта активность, стали Д. А. Поспелов и А. Е. Кибрик. В организации первых встреч активное участие принимал Л. И. Микулич, но вскоре всю тяжесть проведения конференций по проекту «Диалог» взвалили на свои плечи А. С. Нариньяни, Д. Бухштаб и Н. Лауфер, благодаря энергии и труду которых каждый год в Эстонии (как правило, с помощью И. Сильдмяэ под Тарту) или в Подмосковье встречались «стенка на стенку» специалисты по компьютерной лингвистике и искусственному интеллекту. В те годы такие междисциплинарные встречи были уникальным явлением, но именно они, на мой взгляд, и позволили лингвистам и программистам, по сути дела, понять друг друга и проблемы в области создания интеллектуальных диалоговых систем.

Следующей значимой вехой в истории проекта «Диалог» стали, на мой взгляд, семинары «Диалог-2», к участию в которых, в основном, благодаря Д. А. Поспелову, были привлечены когнитивные психологи. К сожалению, такого всплеска междисциплинарных проектов, который обозначился в начале 90-х годов по результатам встреч на семинарах «Диалог-1», не произошло, хотя и «Диалог-2» сыграл свою роль в «наведении мостов» между специалистами по компьютерной лингвистике, когнитивной психологии и искусственному интеллекту.

В настоящее время мы являемся свидетелями (и участниками) нового этапа развития проекта «Диалог», ежегодные конференции которого стали значимой точкой в календаре междисциплинарных конференций по компьютерной лингвистике и интеллектуальным технологиям. В связи с этим, а также в связи с тем, что в качестве «горячей темы» очередной 16-й Международной конференции «Диалог» обозначена тема извлечения информации из ЕЯ-текстов, представляются интересными проведение ретроспективного анализа публикаций в трудах конференций этой серии на данную тему и попытка оценки текущего состояния исследований в данной области в нашей стране.

2. Модели и методы анализа научных сообществ

Следует сразу отметить, что обсуждение исследований и разработок в области анализа научных (и не только научных) сообществ выходит за рамки настоящей работы. Поэтому здесь мы отсылаем заинтересованного читателя к работам[1,2,3,4,5,6,7,8,9] и лишь отметим, что в нашей стране это направление активно развивалось в 70–80-х годах прошлого века. Сейчас эти работы становятся востребованными в связи с практическими задачами выявления «незримых коллективов» и формальной оценки их результативности, как в целом, так и на уровне отдельных членов этих коллективов, а также прогнозирования развития таких сообществ.

[1] *Налимов В. В.* Количественные методы исследования процесса развития науки // Вопросы философии. 1966. № 12. С. 38–47.

[2] *Налимов В. В., Кордон И. В., Корнеева А. Я.* Географическое распределение научной информации // Информационные материалы Научного совета по комплексной проблеме "Кибернетика" АН СССР. 1971. № 2 (49). С. 3–37.

[3] *Nalirnov V. V.* Faces of Science. // Philadelphia, ISI Press, 1981. 297 p.

[4] *Noma E.* Co-citation analysis and the invisible college // J. Amer. Soc. Info. Sci. 1984. Vol. 35. P. 29.

[5] Евстигнеев В. А. Методы теории графов в наукометрии: исследование структуры пространства журналов и незримых коллективов в программировании. // Новосибирск, 1987. (Препр. АН СССР. Новосиб. филиал. ИТМ и ВТ им. С. А. Лебедева; № 4).

[6] *Маршакова И. В.* Система цитирования научной литературы как средство слежения за развитием науки. // М.: Наука, 1988. 288 с,

[7] *Хайтун С. Д.* Проблемы количественного анализа науки. // М.: Наука, 1989. 280 с.

[8] *Паринов С. И.* Контуры единой сетевой инфраструктуры научного сообщества // Технологии информационного общества — Интернет и современное общество: труды V Всероссийской объединенной конференции. СПб., 25–29 ноября 2002 г. СПб.: Изд-во С.-Петерб. ун-та, 2002. С. 118

[9] Егоров В. С., Пожидаев А. В., Чернобровская Т. Н. Систематизация и использование сведений о научных мероприятиях в автоматизированной технологии ВИНИТИ. // НТИ. Сер. 1. — 2006. — №4.– С. 17–23.

3. Статистическая картина мира конференций «Диалог»

3.1. Общий портрет направления

Общая структура исследований и разработок в области извлечения информации из текстов на естественных языках, как она представляется автору на основании анализа публикаций в этой области и собственного опыта, может быть описана рубрикатором, представленным в Табл. 1. Понятно, что данный рубрикатор не претендует на полноту и законченность, но вместе с тем дает основу для оценки ситуации в нашей стране в этой стратегически важной области компьютерной лингвистики и искусственного интеллекта. В связи с этим представляется интересным ретроспективный анализ тех публикаций в трудах конференций серии Диалог, которые (по экспертной оценке автора) относятся к данному направлению. Для проведения такого анализа из трудов конференций за 2000–2009 г. г., представленных на сайте [1], были выбраны работы, которые вписываются в предложенную классификацию. Такое отображение конечно фасетное, так как одна работа (как правило) затрагивает несколько направлений. Полученные результаты представлены в Табл. 1 (наиболее важные, по мнению автора, направления выделены курсивом).

Как показывает анализ представленных выше материалов, из более чем 800 работ, опубликованных в трудах конференций серии Диалог за 2000–2009 г. г., 201 работа (около 25 %) связана тем или иным образом с извлечением информации из текстов. При этом направлению «Модели и методы» посвящены 133 работы (≈ 66.2 %), направлению «Инструментальные средства» — 18 работ (≈ 9 %) и направлению «Интеллектуальные приложения» — 50 работ (≈ 24.8 %). Таким образом, теория «опережает» приложения почти в 3 раза, а инструментарий — почти в 7 раз. На наш взгляд, эти цифры отражают общее состояние исследований в данной области в нашей стране.

Табл. 1.

	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Модели и методы	**6**	**9**	**18**	**11**	**17**	**14**	**13**	**15**	**13**	**17**
• Ресурсы										
• Словари	1		5	1			2	2	3	2
• Тезаурусы	2	2	2	1	2		1	2	2	2
• Онтологии		2	3		1		4	1	1	2
• Подходы к извлечению информации из текстов										
• Статистический	1	1	3	2	4	2	1	2	2	2
• «Легкий» (Shallow)	1	2	1		3	4	3	3	2	3

	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
• «Тяжелый» (Deep)	1	2	4	5	3	4	2	3		2
• ***Гибридный***					1	2		2	1	1
• ***Машинное обучение***										
• ***Автоматическая генерация правил***					1					1
• Онтологический инжиниринг				2	2	2			2	2
Инструментальные средства	**1**	**3**	**4**	**2**	**1**	**1**	**1**	**1**	**0**	**4**
• Средства и языки представления знаний		2	2		1					2
• Платформы	1		2	2		1	1			
• Автоматическое конфигурирование										1
• Отладка		1								1
• ***Тестирование и оценивание результатов***								1		
• Автоматическая генерация процессоров и IE-систем										
Интеллектуальные приложения	**2**	**5**	**6**	**7**	**5**	**7**	**7**	**3**	**4**	**4**
• ***Семантическое аннотирование***		1		1						
• Семантическая кластеризация/классификация		2	4				1		2	3
• Семантический поиск		2	1	1	1	1	1		1	
• Семантическая навигация				1	1					
• Семантическое дайджестирование	1									
• Семантическое реферирование	1				1	2	1			
• Аналитика на знаниях				3	2	3	3	3	1	
• Порталы знаний			1	1		1	1			1
• ***Семантический Веб***										

3.2. Авторский индекс и организации-участники

Для более детальной оценки активности российских специалистов, представленных в конференциях серии Диалог в области извлечения информации из текстов, представляет интерес анализ индекса авторов и организаций этих конференций (Табл. 2).

Табл. 2.

	2000	2001	2002	2003	2004
Специалисты	21	21	43	28	29
Организации	8	11	22	15	17
	2005	**2006**	**2007**	**2000**	**2009**
Специалисты	35	32	36	26	24
Организации	15	15	15	14	18

Как показывает статистический анализ, всего в составе авторских коллективов в трудах конференций Диалог за 2000–2009 г. г. опубликованы работы 295 специалистов из 150 организаций. При этом уникальных авторов — 170, из них представивших более 2 публикаций — 58 чел., более 3 публикаций — 23 чел. и более 4 публикаций — 14 чел.

Наиболее активные авторы конференций серии Диалог по данной тематике представлены в диаграмме на Рис. 1.

Не менее интересен и состав организаций, где, как показывают публикации трудов конференций серии Диалог, ведутся исследования и разработки в области извлечения информации из текстов. Детальный геоландшафт организаций, работающих в данной области, будет представлен в докладе и, частично, дан ниже при анализе семантического портрета конференций серии Диалог. Пока же отметим, что, как и следовало ожидать, наиболее активны в этой области организации Москвы, Новосибирска, Санкт-Петербурга и Казани.

4. Семантическая картина мира конференций «Диалог»

4.1. OntosMiner/SGE — инструментарий анализа

Понятно, что статистические данные, приведенные выше, дают определенное представление о картине мира конференций серии Диалог, но нуждаются в дальнейшей детализации и анализе, который является достаточно трудоемким и потому не может быть проведен без соответствующих инструментальных средств. Детальное обсуждение результатов работы команды Ontos из российской IT-компании

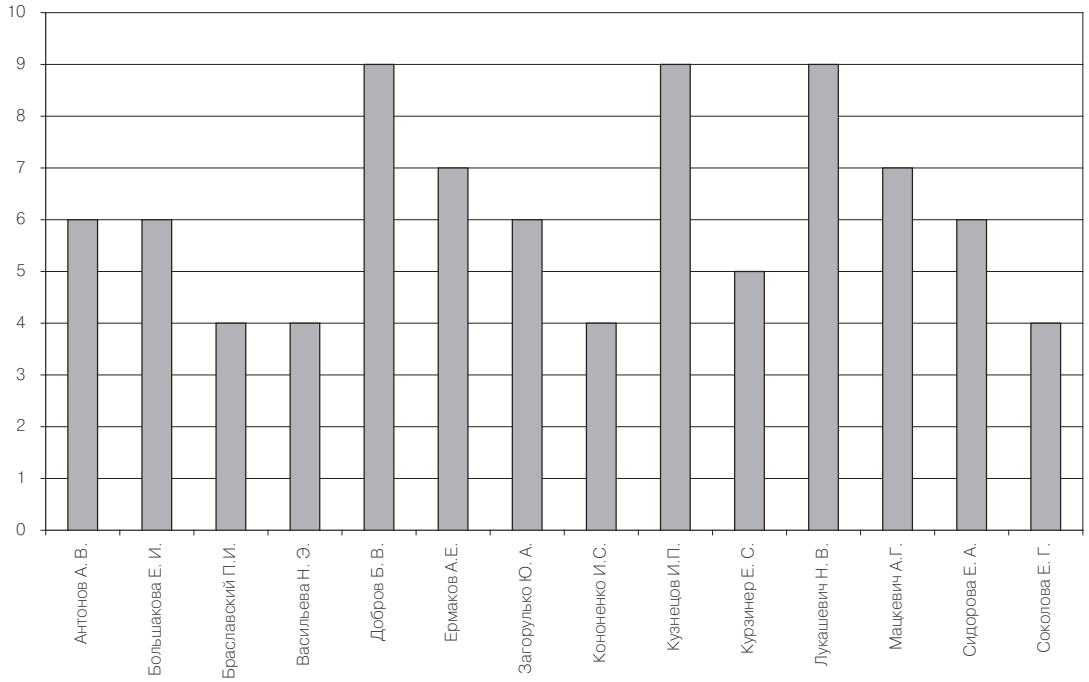


Рис. 1. Наиболее активные авторы конференций Диалог по тематике извлечения информации из текстов

«Авикомп Сервисез» в области извлечения информации из ЕЯ-текстов и соответствующих технологий выходит за рамки настоящей работы. Поэтому ниже кратко описывается только лингвистический процессор OntosMiner/SGE (Shadow Groups Extraction), который использовался в рамках настоящего исследования для построения семантической картины мира конференций Диалог за 2000–2009 г. г.

Концепция разработки системы OntosMiner/SGE лежит в общем русле работ по проекту OntosMiner [2–4] и состоит в следующем:

- Обработка текстов осуществляется под управлением модели предметной области, представленной в виде соответствующей онтологии.
- Для анализа текстов используется shallow approach, в основе которого лежит система шаблонов-образцов на специальном языке представления знаний.
- Результаты обработки отдельных текстов (документов) представляются в виде когнитивных карт (специальных семантических сетей), которые загружаются в базу знаний, где происходит формирование общей семантической сети коллекции документов.
- В качестве инструментальной среды для разработки и реализации системы OntosMiner/SGE использован стандартный инструментарий проекта Ontos, который является развитием среды GATE [5].

В соответствии с общей структурой IE-систем семейства OntosMiner в системе OntosMiner/SGE используется ресурсная цепочка в составе Tokenizer — Morph Tagger — Sentence Splitter — NE Transducer — Semantic Tagger — XML Generator. В качестве первых трех и последней компоненты в системе OntosMiner/SGE используются стандартные ресурсы семейства OntosMiner. Компонента NE Transducer реализуется на основе реинжиниринга соответствующих стандартных модулей, а стандартная компонента Semantic Tagger расширена за счет реализации новых семантических отношений и скомплексирована таким образом, чтобы исключить из нее те стандартные модули, которые обрабатывают семантические отношения, не представленные в данной предметной области.

Общий объем лингвистического процессора OntosMiner/SGE составил 73 специальных и около 600 общих правил, интегрированных из других процессоров семейства, которые компилируются, в соответствии с общей технологией Ontos, в Java-код, а затем и в соответствующую систему Java-классов. Скорость обработки одной статьи стандартного для этих конференций объема (5–9 стр.) составляет, в зависимости от насыщенности ее объектами и отношениями, 1–3 сек на персональном компьютере с процессором Intel® Core™2 Duo с тактовой частотой 2 Ггц и основной памятью 2 Гб.

4.2. Предметная онтология ShadowGroups и корпус текстов

Анализ структуры научных статей, опубликованных в трудах конференций Диалог показал, что с точки зрения целей и задач настоящей работы основными объектами, которые целесообразно извлекать из соответствующих текстов, являются

именованные сущности типа Person, Organization, Location, Paper, Domain, Term, OrgTeam и др., а также семантически значимые отношения между ними типа BeAuthor, BeCoauthor, MemberOf, ReferenceTo, ReferencedBy и т. п. С учетом этого была разработана предметная онтология ShadowGroups Ontology, представленная на Рис. 2.

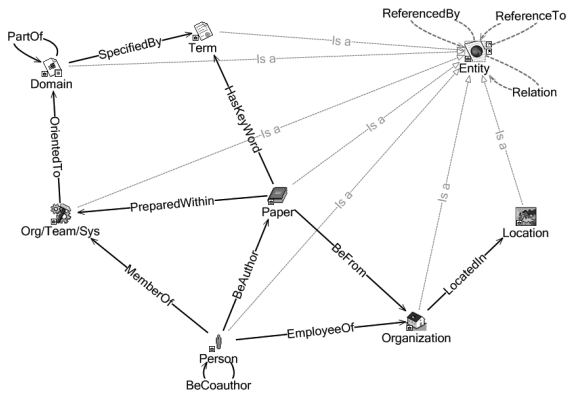


Рис. 2. Онтология предметной области ShadowGroups

Как указывалось выше, в качестве корпуса документов использовались электронные копии трудов конференций Диалог за 2000–2009 г. г., полученные с соответствующего Интернет-сайта. Как правило, в каждом выпуске было около 120 научных статей, покрывающих достаточно широкий спектр исследований и разработок в области теоретической и прикладной лингвистики, а также машинной обработки ЕЯ, но далеко не все они имеют непосредственное отношение к извлечению информации из текстов. Поэтому для дальнейшей обработки из трудов конференций была выбрана 141 статья. Выбранные для дальнейшей обработки статьи были приведены к единому текстовому формату, поскольку на сайте они представлены в форматах txt, doc, pdf и html.

Обработка статей осуществлялась с помощью процессора OntosMiner/SGE под управлением онтологии ShadowGroups, а для объединения результатов в общее семантическое пространство использовались специальные правила идентификации одинаковых объектов, представленных в текстах различным образом. Для визуализации результатов обработки использовалось desktop-приложение LightOntos for Workgroups.

Полученные результаты обсуждаются в оставшихся разделах настоящей работы.

4.3. Ландшафты конференций серии Диалог

Для построения геоландшафта конференций серии Диалог в общей когнитивной карте были оставлены для визуализации только экземпляры понятий Organization и Location. Полученные результаты представлены на Рис. 3.

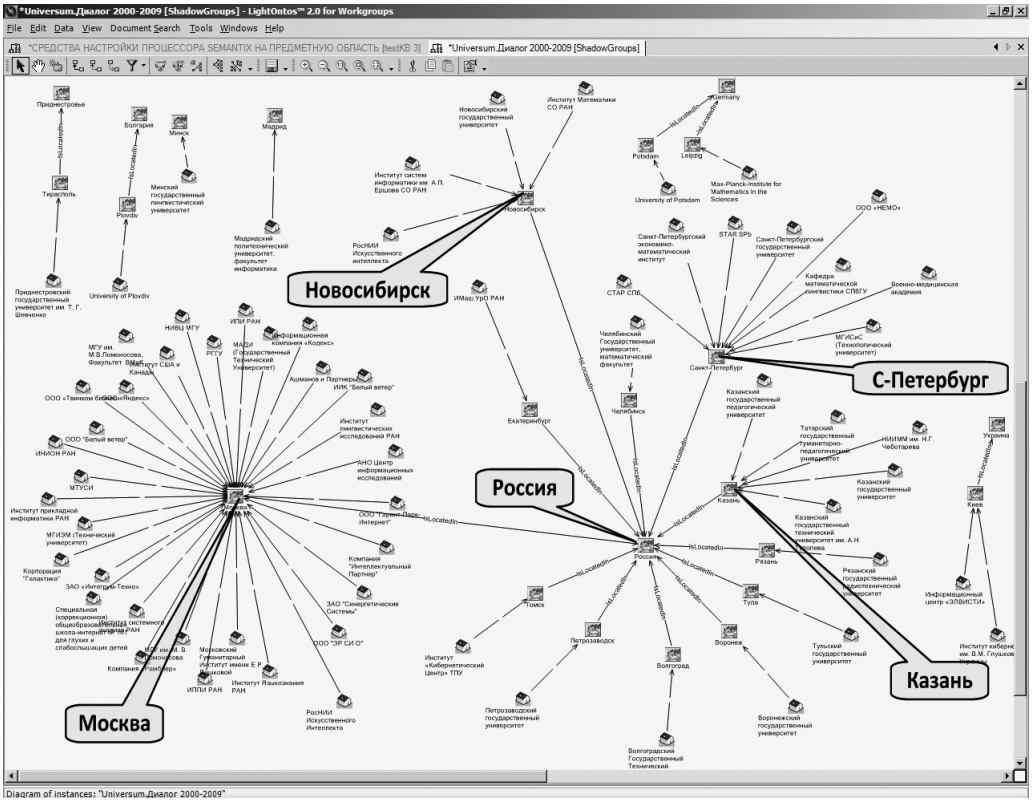


Рис. 3. Геоландшафт конференций серии Диалог

Как показывает анализ этих результатов, наиболее активны на конференциях серии Диалог организации из Москвы (32), Санкт-Петербурга (8), Казани (5) и Новосибирска (4). Из текстов научных статей выделены 26 геообъектов, организации были связаны с геоименами семантическими отношениями LocatedIn, а геоимена с геоименами более высокого уровня (города расположены в странах, если эта информация представлена в статьях) — семантическими отношениями IsLocatedIn. Точность выделения объектов типа Organization составила 83 %, а полнота — 89 %. При этом основные ошибки и пропуски связаны с неполным распознаванием объекта (не всегда выделяется кафедры и факультеты ВУЗов). Точность выделения объектов типа Location составила 93 %, а полнота –96 %. Основные ошибки в определении геоимен связаны, в основном, с ложными срабатываниями правил на фрагментах текстов типа «ПРЕДМЕТНАЯ ОБЛАСТЬ» в названиях статей. Точность и полнота выделения отношений, при условии правильного выделения объектов, — 100 %.

Полученные результаты дают общее представление о функционале процессора OntosMiner/SGE, но с точки зрения целей настоящей работы, на наш взгляд, интереснее обсуждение тематических ландшафтов конференций Диалог разных лет и более детальный анализ авторских коллективов и взаимоотношений между ними.

Всего из корпуса статей, представленных в 2000–2009 г. г. на конференциях Диалог, было выделено объектов типа Person — 1678, Paper — 842, Organization — 204 и Location — 151, между которыми установлены семантические отношеня BeAuthor, BeCoauthor, ReferenceTo и симметричные им отношения ReferencedBy. Характеристики точности и полноты обработки геообъектов даны выше. Точность выделения объектов типа Person составила 91 %, а полнота — 87 %, объектов типа Organization — 83 % и 89 %, а объектов типа Paper — 96 % и 79 %, соответственно. При этом основные ошибки на объектах типа Person были обусловлены необходимостью обработки в рамках одного процессора и русских, и иностранных авторов, использованием модулей из общей библиотеки OntosMiner, которые не адаптировались специально к задаче обработки статей, а также неаккуратностью авторов статей при их оформлении, особенно в коллекциях первых лет. Ошибки на объектах типа Organization, в основном, были связаны с неполнотой общих правил (не идентифицировались факультеты и кафедры как часть организации), а ошибки на объектах типа Paper — разнообразием способов оформления ссылок в коллекциях разных лет. Точность и полнота выделения отношений, при условии правильного выделения необходимых для их сборки объектов, — 100 %.

Как следует из анализа частных диаграмм, конференция Диалог-2000 является наименее насыщенной с точки зрения извлечения информации из текстов. При этом работы разных авторских коллективов, по существу не связаны ни между собой (фокусом когнитивной карты является геообъект «Россия»), ни ссылками на авторитеты, в том числе и западные. Эта же тенденция инкапсуляции ярко выражена в когнитивной карте 2003 года и практически сохраняется в когнитивных картах 2001, 2002 и 2004 г. г. И только в когнитивных картах 2005 г. и, особенно, 2009 г. появляются взаимные ссылки разных авторских коллективов. Авторские ландшафты 2006–2008 г. г. занимают промежуточное положение, поскольку на них видны процессы формирования новых авторских коллективов, с одной стороны, и постепенное появление ссылок между разными авторскими коллективами — с другой.

Вместе с тем, автономный анализ когнитивных карт конференций Диалог разных лет не позволяет понять, существуют ли в области извлечения информации из текстов в нашей стране признанные авторитеты и «скрытые» коллективы. Поэтому в следующем разделе настоящей работы обсуждаются результаты, полученные после объединения всех частных когнитивных карт в единое семантическое пространство.

4.4. Извлечение информации из текстов — скрытые коллективы

Для построения общего семантического пространства конференций серии Диалог отдельные когнитивные карты были обработаны алгоритмами автоматической идентификации одних и тех же объектов и отношений между ними, а полученные результаты верифицировались экспертным путем, а затем проводилось человеко-машинное «схлопывание» оставшихся дублей объектов типа Person, Paper и Organization. Неполнота автоматического объединения объясняется тем, что авторы статей достаточно часто фиксируют в ссылках один и тот же источник различным образом (где-то есть издательство, где-то нет, где-то не указаны страницы, а где-то они есть), названия одних и тех же организаций указываются то полностью, то их аббревиатурами, английские транскрибирования одних и тех же русских, и даже английских, авторов в разных ссылках указываются по разному.

После проведения автоматической идентификации одинаковых объектов в семантическом пространстве осталось объектов типа Person — 919, Paper — 816, Organization — 105 и Location — 25, а после экспертного «выравнивания» — объектов типа Person — 854, Paper — 743, Organization — 71 и Location — 13, а также соответствующие семантические отношения между ними (Рис. 4).

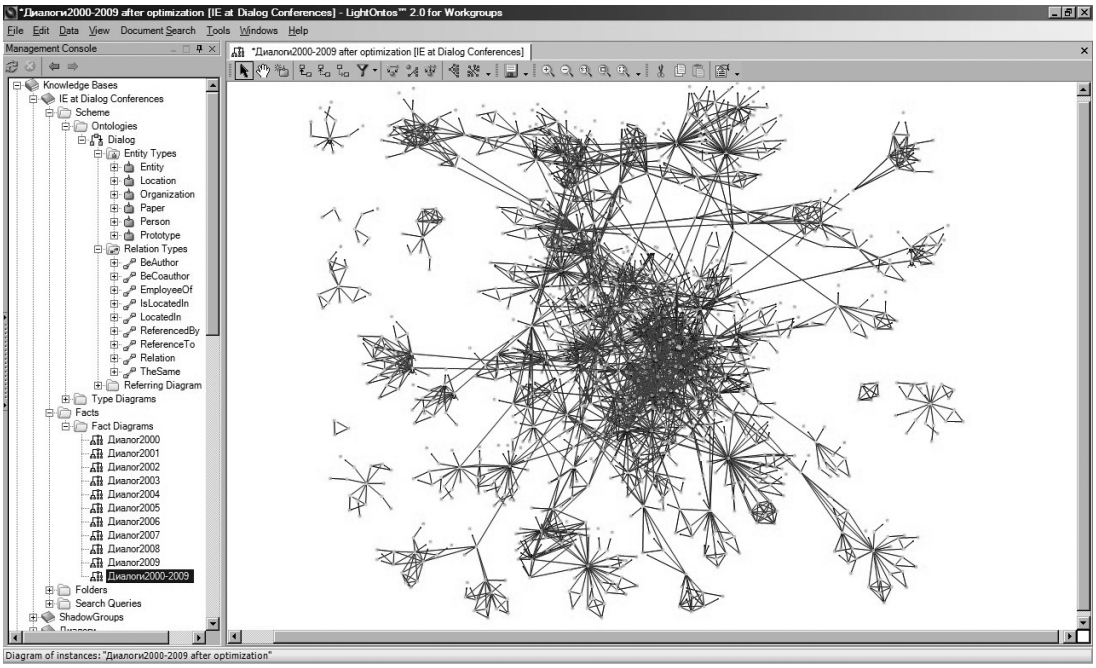


Рис. 4. IE-ландшафт конференций серии ДИАЛОГ

Какие предварительные выводы можно сделать из анализа этой диаграммы? Во-первых, это значительная ее связность при сохранении некоторого числа полностью инкапсулированных авторских коллективов. Во-вторых, появление в диаграмме ярко выраженных и, по-видимому, как-то связанных «скоплений», что позволяет предположить наличие «скрытых коллективов».

Вместе с тем, мелкий масштаб, а также визуализация всех объектов и связей на одной диаграмме не позволяет провести детальный ее анализ. Поэтому для дальнейшего анализа «скроем» на диаграмме все объекты типа Paper, Organization и Location, отношения типа ReferencedBy, поскольку на диаграмме у них есть симметричные отношения типа ReferenceTo а также удалим изолированные подграфы. Диаграмма, полученная в результате таких преобразований, представлена на Рис. 5 и отражает возможные коллективы и связи между ними.

Для дальнейшего упрощения диаграммы, представленной на Рис. 5, уберем с карты все «желтые» авторские коллективы, тех, кто не попал в «подсветку», а также тех авторов, на которых ссылаются специалисты только одной команды. В результате получим диаграмму, представленную на Рис. 6.

Анализ персоналий потенциальных авторитетов в области IE по данным конференций серии Диалог показывает, что их можно разделить на три категории. В первой — общепризнанные авторитеты в области лингвистики, которые практически не работают в области извлечения информации из текстов и потому ссылки на них можно назвать «ритуальными» [6–10 и др.]. Во второй — ссылки на учебники и фундаментальные работы, которые можно отнести к образовательным [11–14 и др.]. И, наконец, в третьей категории — работы, которые повлияли на текущие исследования коллективов нашей страны в данной области в теоретическом и/или практическом плане [15–18 и др.].

Не менее, а может и более интересной является диаграмма, представленная на Рис. 7, где детализируется взаимодействие между разными авторскими коллективами и внутри каждого из них.

Как показывает анализ этой диаграммы, все «команды» в значительной мере «замкнуты» на себя и при этом подавляющая часть авторов «грешат» автоссылками. Самая многочисленная «команда» представляет несколько тесно связанных между собой организаций и коллективов из Новосибирска (кластер Загорулько [19,20 и др.]), который, при этом, никак не связан с другой новосибирской командой (кластер Загоруйко [21]), которая ссылается только на кластер Кузнецова (ИПИ РАН) [22,23]. Кластер Загорулько ссылается на кластер Ермакова, например, на [24], на кластер Добров-Лукашевич [25] (АНО Центр информационных исследований), например, на [25–27], который, в свою очередь, тесно связан общими исследованиями и статьями [28] с кластером Невзоровой (Казань), а взаимными ссылками — с кластером Ермакова (RCO) [24,28], где есть ссылка на кластер Антонова (корпорация «Галактика») [30,31] и кластер Кузнецова [22], а из кластера Кузнецова на кластер Ермакова [24]. Из диаграммы на Рис. 7 следует, что наиболее «открыт» миру IE-кластеров нашей страны кластер Большаковой [32,33 и др.] (ВМиК МГУ), где имеются несколько внешних ссылок [26, 27 и др.]. И, наконец, единственный кластер, на который ссылаются три других кластера — это кластер Браславского (ИМаш УрО РАН) и работы [34,35].

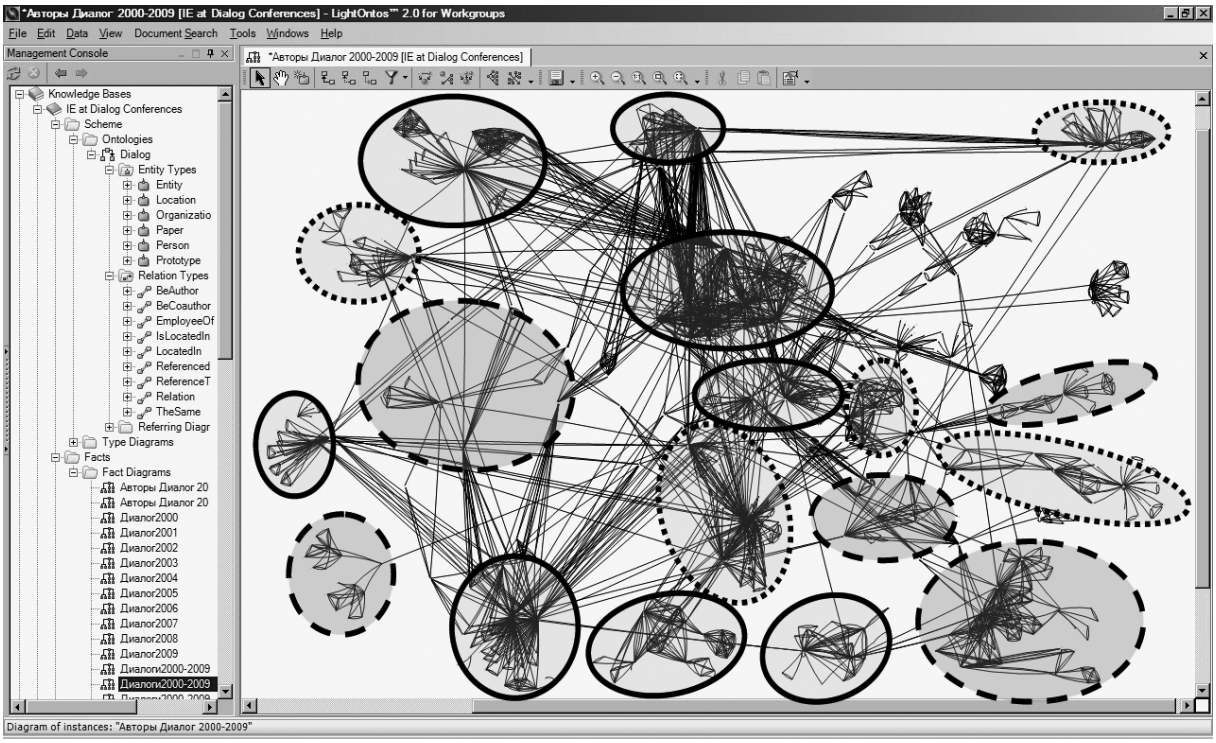


Рис. 5. Диаграмма связей авторских коллективов в области IE[10]

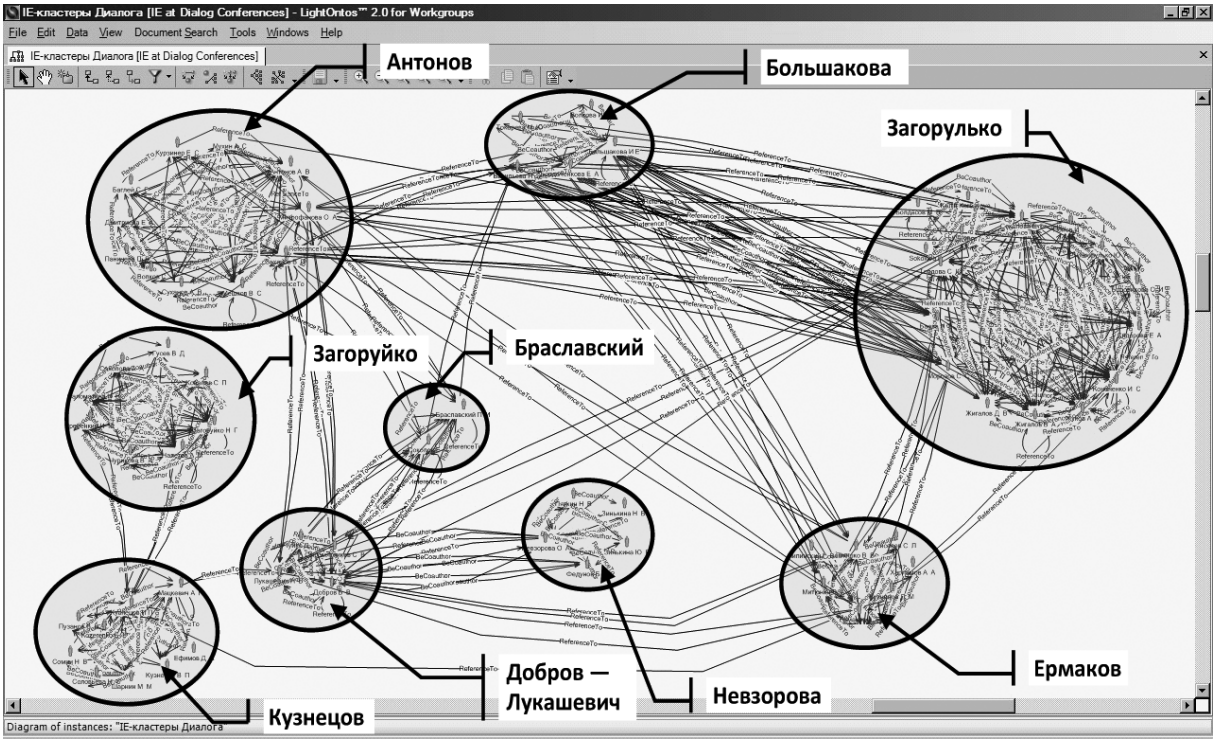


Рис. 6. Диаграмма потенциальных авторитетов и активных IE-кластеров[11]

[10] Точечной обводкой выделены известные коллективы, которые, как представляется, занимаются смежными и/или частными, по отношению к общей проблеме извлечения информации из текстов, исследованиями, пунктирной — потенциальные фокусы «скрытых» коллективов (авторитеты в области IE), а сплошной — коллективы, активно работающие в этой области.

[11] Пунктирной обводкой, как и раньше, выделены потенциальные авторитеты, а сплошной — наиболее активные кластеры, которые для удобства дальнейшего обсуждения связаны с фамилиями их лидеров.

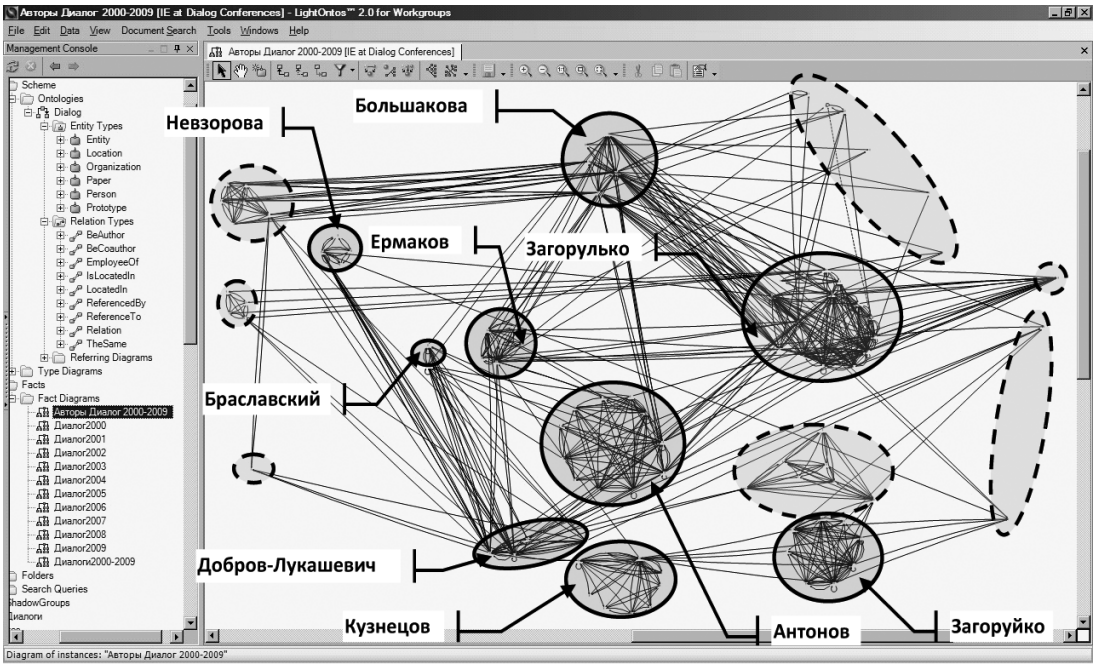


Рис. 7. Диаграмма взаимодействия активных IE-кластеров

Анализ перечисленных выше кластеров и работ российских специалистов, в них представленных, показывает, что в конференциях серии Диалог представлены только 3 коллектива (ИПИ РАН, RCO и ВМиК МГУ), основная деятельность которых связана с разработкой и реализацией систем типа IE, и 2 коллектива (АНО Центр информационных исследований и РосНИИ искусственного интеллекта), которые, скорее, можно отнести к области онтологического инжиниринга и, в частности, использования его результатов для извлечения информации из текстов. Подробное обсуждение исследований и разработок, которые ведутся в этих коллективах выходит за рамки настоящей статьи, но в целом можно констатировать, фронт этих работ явно недостаточен для решения такой важной и сложной проблемы как извлечение информации из текстов.

5. Несколько замечаний вместо заключения

В настоящей работе дан ретроспективный анализ публикаций конференций серии Диалог за 2000–2009 г. г. в области извлечения информации из текстов. Представлены статистическая и семантическая картины публикаций этих конференций в области извлечения информации из текстов.

Как показал проведенный анализ, в рамках конференций серии Диалог исследованиям и разработкам в этой области всегда уделялось определенное внимание. Вместе с тем, детальный анализ публикаций позволяет высказать несколько замечаний по поводу ситуации, которая сложилась в рамках исследований и разработок по извлечению знаний из ЕЯ-текстов в нашей стране:

1. Критическая масса специалистов и организаций, активно работающих в этой области, явно недостаточна по сравнению с общемировыми тенденциями, а география исследований и разработок ограничена традиционными центрами.
2. Теоретические исследования интересуют российских специалистов существенно больше, чем использование их результатов для создания практически значимых систем.
3. Практические разработки, как правило, доводятся лишь до уровня прототипов, а в публикациях отражаются лишь их отдельные аспекты, которые не дают возможности оценить уровень соответствующих систем.
4. Российские авторы публикаций в этой области лучше знают работы зарубежных коллективов, чем отечественных.
5. Практически все российские авторские коллективы «грешат» автоссылками, что, в конечном счете, показывает отсутствие признанных авторитетов и, как следствие, «скрытых коллективов» в этой области.

Учитывая все вышесказанное, представляется, что для российских специалистов в области извлечения информации из текстов время полномасштабного и скоординированного развертывания исследований и разработок в этой важнейшей области компьютерной лингвистики и интеллектуальных технологий еще впереди. И конференции проекта Диалог, как и конференции других смежных направлений, могут и должны здесь сыграть важную роль.

Литература

1. *Труды* международных конференций по компьютерной лингвистике и интеллектуальным технологиям: «Диалог 2000» — «Диалог 2009». http://www.dialog-21.ru
2. *Хорошевский В. Ф.* OntosMiner: семейство систем извлечения информации из мультиязычных коллекций документов. // Труды конференции КИИ-2004, Тверь, Россия, 2004.
3. *Минор С. А., Старостин А. С.* Ontos: технология извлечения знаний из неструктурированных текстов и семантическое индексирование. // Доклад на международной конференции по компьютерной лингвистике и интеллектуальным технологиям «Диалог 2007». (Бекасово, 30 мая — 3 июня 2007 г.).
4. *Efimenko I., Hladky D., Khoroshevsky V., Klintsov V.* Semantic Technologies and Information Integration: Semantic Wine in Media Wine-skin. // Proc. of the 2nd European Semantic Technology Conference (ESTC2008), Vienna, 2008.
5. *Cunningham H., Maynard D., Bontcheva K., Tablan V.* GATE: an Architecture for Development of Robust HLT Applications // Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL), Philadelphia, July 2002.
6. *Апресян Ю. Д.* Избранные труды, том II . Интегральное описание языка и системная лексикография. // М., 1995.
7. *Иорданская Л.* (1992) Коммуникативная структура и ее использование в системе текстовой генерации // Международный форум по информации и документации. Т. 17, № 2.
8. *Мельчук И. А.* Опыт теории лингвистических моделей «Смысл — Текст» // М .: Наука ,1974 .
9. *Нариньяни А. С.* Автоматическое понимание текста — новая перспектива // Труды международного семинара Диалог'97 по компьютерной лингвистике и ее приложениям. — Москва, 1997.
10. *Ершов А. П.* К методологии построения диалоговых систем: феномен деловой прозы // Избранные труды. Новосибирск: ВО «Наука», 1994.
11. *Gruber T. R.* A translation approach to portable ontologies // Knowledge Acquisition, 1993, V. 5(2), P. 199–220
12. *Sowa, J. F.* Knowledge Representation: Logical, Philosophical, and Computational Foundations, Brooks Cole Publishing Co., Pacific Grove, CA. 2000 .
13. *Вежбицка А.* Метатекст в тексте // Новое в зарубежной лингвистике. Вып. VIII . М.: Прогресс, 1978.
14. *Гаврилова Т. А., Хорошевский В. Ф.* Базы знаний интеллектуальных систем. Учебник. // СПб.:Питер, 2001.
15. *Lin D.* Using syntactic dependency as local context to resolve word sense ambiguity // Proceedings of the 35th annual meeting on Association for Computational Linguistics. Madrid , Spain , 1997.
16. *Зализняк А. А.* Грамматический словарь русского языка. Словоизменение. // М., "Русский язык", 1977.
17. *Большаков И. А.* Многофункциональный словарь-тезаурус для автоматизированной подготовки русских текстов // НТИ сер. 2. 1994. — N 1.
18. *Сокирко А. В.* Морфологические модули на сайте www.aot.ru // Компьютерная лингвистика и интеллектуальные технологии: Тр. междунар. конференции Диалог'2004 («Верхневолжский», 2–7 июня 2004 г.) / Под ред. И. М. Кобозевой, А. С. Нариньяни, В. П. Селегея. М.: Наука, 2004.
19. *Kononenko I., Kononenko S., Popov I., Zagorul'ko Yu.* Information Extraction from Non-Segmented Text (on the material of weather forecast telegrams). // Content-Based Multimedia Information Access. RIAO'2000 Conference Proceedings, v.2, 2000.
20. *Сидорова Е. А.* Многоцелевая словарная подсистема извлечения предметной лексики // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог–2008». М.: 2008.
21. *Загоруйко Н. Г., Налетов А. М., Гребенкин И. М.* На пути к автоматическому построению онтологии. // Труды конференции Диалог-2003.
22. *Кузнецов И. П.* Семантические представления. // М. Наука. 1986. 290 с.
23. *Кузнецов В. П., Мацкевич А. Г.* Автоматическое выявление из документов значимой информации с помощью шаблонных слов и контекста. // Труды международного семинара Диалог-98 по компьютерной лингвистике и ее приложениям. Том 2. Казань 1998.
24. *Ермаков А. Е.* Автоматическое извлечение фактов из текстов досье: опыт установления анафорических связей // Труды международной конференции Диалог'2007 «Компьютерная лингвистика и интеллектуальные технологии». М.: Наука, 2007.
25. *Добров Б. В., Лукашевич Н. В.* Онтологии для автоматической обработки текстов: Описание понятий и лексических значений // Компьютерная лингвистика и интеллектуальные технологии: Тр. междунар. конференции Диалог'06, Бекасово, 31 мая — 4 июня 2006 г., 2006.
26. *Добров Б. В., Лукашевич Н. В., Сыромятников С. В.* Формирование базы терминологических словосочетаний по текстам предметной области // Труды пятой всероссийской научной конференции «Электронные библиотеки: Перспективные методы и технологии, электронные коллекции», 2003.

27. *Лукашевич Н. В., Добров Б. В.* Тезаурус русского языка для автоматической обработки больших текстовых коллекций // Компьютерная лингвистика и интеллектуальные технологии: Труды Международного семинара Диалог'2002 — М.: Наука, 2002.
28. *Добров Б. В., Лукашевич Н. В., Невзорова О. А., Федунов Б. Е.* Методы и средства автоматизированного проектирования прикладной онтологии // Известия РАН. Теория и системы управления. М.: 2004. № 6.
29. *Ермаков А. Е., Плешко В. В., Митюнин В. А.* RCO Pattern Extractor : компонент выделения особых объектов в тексте // Информатизация и информационная безопасность правоохранительных органов: XI Международная научная конференция. Сборник трудов. М.: 2003.
30. *Антонов А., Курзинер Е.* Автоматическое выделение предметной области большого необработанного текстового массива // Компьютерная лингвистика и интеллектуальные технологии, Труды Международного семинара Диалог-2002.
31. *Баглей С. Г., Антонов А. В., Мешков В. С., Суханов А. В.* Кластеризация документов с использованием метаинформации // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог-2006». М.: 2006.
32. *Большакова Е. И., Баева Н. В., Бордаченкова Е. А., Васильева Н. Э., Морозов С. С.* Лексико-синтаксические шаблоны в задачах автоматической обработки текста // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог–2007». М.: 2007.
33. *Большакова Е. И., Васильева Н. Э., Морозов С. С.* Лексико-синтаксические шаблоны для автоматического анализа научно-технических текстов // Десятая Национальная конференция по искусственному интеллекту с международным участием КИИ-2006. Труды конференции в 3-х томах. М.: Физматлит, 2006. Браславский П. И., Соколов Е. А. Сравнение четырех методов автоматического извлечения двухсловных терминов из текста // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог–2006». М.: 2006.
34. *Браславский П. И., Соколов Е. А.* Автоматическое извлечение терминологии с использованием поисковых машин интернета // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог–2007». М.: 2007.
35. *Браславский П. И., Соколов Е. А.* Сравнение пяти методов извлечения терминов произвольной длины // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог–2008». М.: 2008.

Именные предикативы и дативные предложения в европейских языках[1]

Nominal predicatives and syntactic structures with a dative marking on the semantic subject in the european languages

Циммерлинг А. В. (meinmat@yahoo.com)

Московский государственный педагогический университет

В статье обсуждается соотношение между лексической группировкой несогласуемых именных слов, выражающих предикатное значение Состояния, и моделями предложения, где семантический субъект маркируется дат. п. В тех европейских языках, где имеются именные предикативы, как правило, есть и дативные предложения, но между этими двумя явлениями нет безусловной связи. Для типологии предикативов ключевое значение имеет противопоставление классов адъективных основ: от основ одного класса образуются только обозначения постоянных свойств, от основ другого класса — только слова ситуативного признака, основы третьего класса амбивалентны. Соотношение этих трех классов основ в конкретном языке определяет число предикативов и выражаемые ими лексические значения.

Опираясь на гипотезу Л. В. Щербы о наличии в русском языке Категории Состояния как особой части речи, В. В. Виноградов (1947), Н. С. Поспелов (1955) и А. В. Исаченко (1955) связали наличие группировки именных предикатов состояния с особенностями русского морфосинтаксиса — грамматикализацией нулевой связки «Быть» в 3 л. ед. ч. наст.времени изъяв. накл. и широким распространением дативно-безличного предложения, где семантический субъект маркируется дат. п., а в позиции предиката стоит неглагольное слово, выражающее идею пребывания конкретного одушевленного субъекта в определенном психологическом и/или физиологическом состоянии в течение некоторого отрезка времени (ср. *X-у мутно*, *стыдно*, *грустно, весело, жаль* и т. п.).

Валентность на дат. п. лица у русских предикативов не является тривиальным свойством — она очерчивает семантический класс внутреннего состояния. Вслед за (Булыгина 1982), (Селиверстова 1982), определим предикаты этого класса как выражения, указывающие на пребывание субъекта в некотором неизменном состоянии, не являющемся результатом чьего-либо непосредственного воздействия[2], в течение некоторого отрезка времени[3]. Такому определению в русском языке удовлетворя-

---

[1] Работа выполнена при поддержке гранта РГНФ 09-04-00297а «Типология синтаксических ограничений».

[2] По этому признаку предикаты состояния противопоставлены т. н. результативам, т. е. предикатам, указывающим, что пребывание лица или предмета в данном состоянии есть следствие некоторых процессуальных изменений. Стандартными результативами в русском языке являются предикатные формы причастий, соотнесенные с исходными глагольными предложениями: так рус. *ручка сломана* соотносится с ситуацией, обозначаемой предложением *X сломал ручку*, а рус. *Вася рассержен* может соотноситься как с ситуацией, обозначаемой предложением *X рассердил Васю*, так и с ситуацией, обозначаемой предложением *Вася рассердился*. В русском языке имеется безличный результатив с формой причастия: *мне разрешено/запрещено рисовать*. Такие предложения соотнесены с глагольными предикатами вида *X разрешил/запретил Y-у Z*, поэтому они не могут быть признаны обозначениями состояний. Напротив, предложения *мне положено/не положено рисовать* могут быть признаны обозначениями состояний, поскольку на синхронном уровне они не соотносятся с глагольными предложениями **X положил Y-у (не) рисовать*.

[3] Этот признак, как показал уже Л. В. Щерба (Щерба 1974) противопоставляет состояния предикатам свойства (качества), выражающих вневременные характеристики, ср. *Катя — студентка*, *Катя красивая/ Катя — красивая женщина* и т. п.

ет ок. 150–160 лексем, способных формировать дативное предложение (*X-у* ***пора*** *уходить, X-у* ***стыдно*** *за свое поведение, X-у* ***грустно****, что его дочь опоздала*); если же считать и устойчивые сочетания вроде *X-у* ***не с руки****, X-у* ***не к спеху***, число увеличится еще на сотню. Подавляющее большинство русских слов состояния — производные формы. Наличие у предикатива коррелята в виде прилагательного (*грустный* vs **стыдный*) нерелевантно для предикации состояния в русском языке, но несогласуемые формы с финалью *-о*, способные употребляться в структуре N_{dat} — V_{link} — Pred[4], образуются не от всех адъективных основ. То, что способность формировать дативное предложение есть семантический тест, выявляется не при рассмотрении изолятов типа *стыдно, совестно, жаль, надо, пора*, которые лишены согласовательных коррелятов **стыдный, *совестный*, а при рассмотрении тех слов с финалью-*о*, которые соотнесены с прилагательными. Не все предикативы на *-о,* которые употребляться в схемах Adv — Pred и *Это*— Pred, имеют валентность на дат. п. лица: ср. правильные предложения *здесь пыльно, сегодня морозно, это алогично, это аморально* при аномальных **мне пыльно, *мне морозно, *вам алогично, *все аморально*. Такое распределение подтверждает, что предикативы с валентностью на дат. п. лица действительно имеют таксономические признаки состояний. В отличие от Н. С. Поспелова, мы не настаиваем, что те предикативы, которые возможны в дативном предложении, являются единственными обозначениями состояний в русском языке, и допускаем, что под то же таксономическое значение можно подвести несогласуемые предикаты вроде *X навеселе/наготове/не в своей тарелке*, реализующиеся в двусоставной схеме при подлежащем в им. п.

В работах (Бонч-Осмоловская 2003) и (Гришина 2002), опирающихся на (Грамматика 1980), утверждается, что в русском языке имеется единое дативное предложение с присущими ему инвариантными структурно-семантическими свойствами. Этот тезис не подтверждается анализом — между маркировкой семантического субъекта поверхностным дат. п. и семантикой предложения нет явной связи. Два типа дативных схем — Дативно-Инфинитивные структуры (ДИС), ср. *грузовикам здесь не проехать, нам еще тренироваться и тренироваться, всем отправляться по домам* и Дативно-Предикативные структуры (ДПС), ср. *мне тошно, стыдно, мне противно это говорить, мне грустно, что она опаздывает* имеют в русском языке разные структурные и семантические свойства, что доказывается в (Zimmerling 2009), (Циммерлинг 2009) и иллюстрируется на рис. 1 ниже.

[4] От несогласуемых предикатов состояния с финалью -о необходимо отличать согласуемые формы прилагательных, которые тоже могут иметь валентность на дат. п. лица. Ср. предикатив *противно* в предложениях рус. *Мне противно читать такое*, *Мне противно, что журналисты сеют сплетни* и краткое прилагательное ср. р. *чуждо* в предложении *Мне это чуждо* (контрольное предложение *мне* [*эти устремления* $_{им.\ п.\ мн.\ ч.}$ *чужды* $_{им.\ п.\ мн.\ ч.}$]). Валентность на дат. п. лица может быть присуща и согласуемым формам тех кратких прилагательных, от которых образуются предикативы, ср. *мне* [*противны* $_{им.\ п.\ мн.\ ч.}$ *его поползновения* $_{им.\ п.\ мн.\ ч.}$], но форму *чуждо* нельзя употребить в синтаксических контекстах вроде **мне чуждо, что журналисты сеют сплетни*, ??*мне чуждо повторять сплетни*.

	ДИС	ДПС
Минимальная схема	[N_{dat} — V_{inf}]	[N_{dat} — V_{link} — Pred]
Расширенная реализация	[N_{dat} — Pred— V_{inf}]	[N_{dat} — V_{link} — Pred] — V_{inf}
{+ Одушевленность} субъекта	Нет	Да
Семантическая роль субъекта (базовое значение)	Модальный субъект (алетическая модальность)	Субъект состояния
Интерференция с сентенциальными подлежащими (CPs)	Нет	Да
Лексический охват	Открытый класс	150–160 именных предикативов
Механизм приписывания дат. п. семантического субъекта	Нелексический: дат. п. не приписывается какой-либо вершиной и является встроенной характеристикой нефинитного предложения	Лексический: дат. п. приписывается предикативом как лексической вершиной

Рис. 1. Параметры ДИС и ДПС в русском языке

Тем не менее, есть языки, где значения, выражаемые в русском языке с помощью ДИС и ДПС, выражаются на основе одной и той же синтаксической схемы. Так, например, в латышском языке предикативов, образованных от основ прилагательных, нет, но есть особая глагольная форма т. н. дебитива, которая требует оформления семантического субъекта дат. п. Модальное значение «внешнего», т. н. алетического, долженствования, выражается добавлением приставки jā- к финитной форме 3 л. глагола.

(7) *Man nebūtu jāstrādā, ja es būtu bagāts.*
'Мне не надо было бы работать, если бы я был богат'.

Возможны две точки зрения на грамматический статус латышского дебитива. А. Хольфут отвергает тезис о том, что дебитив — особое наклонение латышского глагола, и показывает, что финитная форма с приставкой jā- сочетается с показателями индикатива, конъюнктива и прочих наклонений, выражаемых формами вспомогательного (выполняющего роль связки) глагола *būt* 'быть' (Holvoet 2001, 32–46). Сам А. Хольфут трактует дебитив как финитную модальную форму латышского глагола, семантически эквивалентную сочетаниям модального глагола с инфинитивом (ср. рус. *надо/предстоит работать, надо было бы работать, надо будет работать*). А. В. Андронов отталкивается от того факта, что латышский дебитив реализуется в связочных структурах[5], и видит в нем аналог латинского герундива или русских предикативных прилагательных (Andronov 1999), т. е. предикативов в терминах данной статьи. Так или иначе, употребление дебитива частично покрывает ту зону, которая в русском языке покрывается употреблением именных слов типа *надо, нужно*, а также модальных глаголов *предстоит, следует, надлежит*. В то же время, в латышском языке нет ограничения на одушевленность семантического субъекта предложения с дебитивом, ср. пример (5). Это сближает предложения с формой дебитива с русскими дативно-инфинитивными структурами (ДИС). Кроме того, дебитив, как и ДИС в русском языке, может образовываться от любого глагола, независимо от его семантики.

(8) *Es*$_{им.\ п.}$ *lasu grāmatu*$_{вин.п}$.
«Я читаю книгу».

(9) *Grāmata*$_{им.\ п.}$ *ir lasīta*$_{им.\ п.}$.
Букв. «Книга есть прочтена».

(10) *Man*$_{дат.\ п.}$ *(ir) jālasa grāmata*$_{им.\ п.}$.
«Мне [надо] прочесть книгу», букв. «мне надо-читает книга».

(11) *Grāmatai*$_{дат.\ п.}$ *ir jābūt lasītai*$_{дат.}$ п.
«Книга должна быть прочтена»,
букв. «Книге [связка] надо-есть прочитанной».

(12) *Man*$_{дат.\ п.}$ *bija jālasa grāmata*$_{им.\ п.}$.
«Мне надо было прочесть книгу»,
букв «мне было надо-читает книга».

(13) *Man*$_{дат.\ п.}$ *būtu jālasa grāmata*$_{им.\ п.}$.
«Мне надо будет прочесть книгу»,
букв «мне будет надо-читает книга».

(14) *Viņai*$_{дат.\ п.\ ж.\ р.}$ *būšot jālasa grāmata*$_{им.\ п.}$.
«Ей надо было бы прочесть книгу»,
букв «мне было-бы надо-читает книга».

[5] В том числе, в структурах с нулевой связкой/опущением связки 3 л. ir в наст.времени индикатива.

Семантический субъект латышского дебитива обладает рядом свойств подлежащего (не в меньшей степени, чем субъект ДИС в русском языке[6]), см. примеры (9–11), иллюстрирующие способность формы дат. п. man «мне» контролировать употребление возвратного местоимения, и (12), где показывается способность ИГ Inetei контролировать согласование со именной частью сказуемого (ср. *рус. Ей нравится жить* ***одной***, *Васе нравится решать задачи* ***самому***).

(15) *Man*$_{дат.\ п.}$ *jānopērk un jāizlasa šī*$_i$ *avīze*$_{им.\ п.}$
«Мне [надо] купить и читать себе газету»,
«Мне$_i$ — покупать и читать себе$_i$ газету», букв.
«мне$_i$ надо-читает и надо-покупает себе$_i$ газета».

(16) *(Man*$_i$*)*$_{дат.\ п.}$*jāizlasa avīze*$_{им.\ п.}$*. [par sevi]*$_i$.
«(Мне$_i$) [надо] читать газету про себя$_i$»,
букв. «мне$_i$ надо-читает газета про себя$_i$».

(17) *(Man*$_i$*)*$_{дат.\ п.}$ *[pašai]*$_i$ *jāizlasa avīze*$_{им.\ п.}$.
«(Мне)$_i$ самому$_i$ читать газету», букв.
«(Мне)$_i$ самому$_i$ надо-читает газету».

(18) *[Inetei*$_i$*]*$_{дат.\ п.\ ж.\ р.\ ед.\ ч.}$*jābūt [labai*$_i$*]*$_{дат.\ п.\ ж.\ р.\ ед.\ ч.}$.
«Инете$_i$ должна быть милой/доброй», «Инете надо быть милой$_i$».

Как и русские предложения с операторными словами *должен* и *надо*, латышские предложения с формой дебитива наст.времени типа (12) неоднозначны: они могут выражать как значение деонтической модальности — 1) «Если Инете соблюдает определенную норму или конвенцию, она должна быть милой, ее долг — быть милой», так и значение алетической модальности — 2) «Внешние обстоятельства, ход событий вынуждают Инете быть милой, у нее нет выбора». Напротив, внешне аналогичные русские предложения с ДИС, по-видимому, не могут выражать деонтические значения и всегда выражают алетическую модальность, т. е. концептуализуют ситуацию внешнего принуждения. В этом плане предложения (13) и (14), не являются контрастной парой, хотя в первом случае ИГ в дат. п. одушевленная, а во-втором — нет.

(19) *Грузовикам тут не не проехать.*
(Ход событий полностью исключает Р).

(20) *Девушкам — идти в постель и улыбаться!*
<Такова воля Атамана.>
(Внешнее принуждение делает Р неизбежным[7]).

[6] Синтаксические свойства подлежащеподобных актантов в русском языке разбираются в [Zimmerling 2009].

[7] Точнее, говорящий, не считаясь с волей адресата, приказывает адресату избрать сценарий поведения, при котором Р становится неизбежным. Речевой акт приказа приписывает говорящему способность добиться осуществления Р помимо воли адресата.

Напротив, высказывания с *надо* и *нужно* во многих контекстах допускают (15) или даже требуют (16) деонтическую интерпретацию.

(21) *Девушкам* ***надо*** *почаще улыбаться.*
<Тогда их карьера и личная жизнь будут успешными>.

(22) *Инессе* ***не нужно*** *улыбаться во время доклада начальника. <Это действие может повредить X-у>.*

Итак, семантическое противопоставление модальных предикатов в конструкции, где ядерным компонентом является инфинитив (ДИС) и в конструкции, где ядерным компонентом является предикатив, а инфинитив, если он вообще допустим, является нисходящей валентностью предикатива (ДПС), сводится к формуле:

- ДПС используется в русском языке и для предикации деонтической, и для предикации алетической модальности, но ДИС не используется в русском языке для выражения деонтических значений и всегда выражает алетическую модальность, значение внешнего принуждения.

Противопоставление модальных предикатов по признаку способности/неспособности выступать в качестве деонтических операторов является еще одним аргументом против объединения ДПС, ср. (14) и ДИС, ср. (15), (16) в рамках единой конструкции с якобы присущей ей общей семантикой. Для латышского языка, где предложения в дебитиве могут выражать оба типа модальных значений, гипотеза о наличии единой предикатной семантики у безличных схем с семантическим субъектом в дат. п., уместнее. Напомним, что в латышском языке, в отличие от русского, нет именных предикативов и отсутствует ДПС.

ДПС с параметрами, близкими к русским, представлено в ряде славянских и германских языков. Естественно проверить гипотезу о том, что в языках, где есть дативное предложение, есть и лексическая группировка предикативов, в том числе, слова, лишенные непредикативных коррелятов и употребляющиеся только в позиции именного сказуемого при связочном глаголе (ср. рус. *жаль, стыдно, надо*). Данная гипотеза, высказанная в (Zimmerling 1992) подтвердилась материалом: в древнескандинавских памятников нашлось около 1000 предикативов, ср. глоссарий в (Циммерлинг 2002). Сравним ситуацию в четырех языках с близкими морфосинтаксическими характеристиками — современном русском, древнеисландском, новоисландском и современном чешском. Вначале оценим ДПС в русском. Общее число предикативов со значением состояния составляет 250–260 единиц, из них 150–160 образованы от адъективных основ: пополнение класса возможно, но новые предикативы являются близкими синонимами уже существующих — ср. разг. *X-у фиолетово, лениво* и лит. Рус. *X-у все равно, лень*. В русском языке предикат состояния всегда имеет конкретного одушевленного субъекта и соотнесен со значением актуального ситуативного признака: высказывания типа **быть тошно после Нового Года случается нередко* аномальны. То, что этот запрет нетривиален, показывает новоисландский язык, где предложения типа (17b), с инфинитивной группой и предикативом в позиции подлежащего, грамматичны.

(17a)
Honum *varð* *flökurt*.
Он$_{\text{дат. п. ед.ч. мр.}}$ стать$_{\text{3 л. ед. ч. прш. вр.}}$ тошно$_{\text{ср. р. ед. ч.}}$
‘Ему стало тошно’.

(17b)
[Að verða flökurt] *er*
Чтобы становиться$_{\text{инф.}}$ тошно$_{\text{ср. ед. ч.}}$ быть$_{\text{3 л. ед. ч.}}$
algengt.
повсеместный$_{\text{ср. р. ед. ч.}}$
‘Люди часто испытывают тошноту’,
Букв. ‘[быть тошно] случается часто’.

В (17a) предикатив *flökurt* ‘тошно’, как и его русский эквивалент, актуализован и описывает референтную ситуацию, где конкретный X испытывает тошноту. И новоисландский, и русский предикативы соотнесены с адъективными парадигмами и омонимичны согласуемым прилагательным в форме им.-вин. п. ср. р. ед. ч. Согласуемых форм **тошный*, **flökurr* ни в том, ни в другом языке нет, т. е. с точки исландца и русского, не бывает конкретных референтов (предметов или лиц), к которым можно приложить признак «быть тошным». Тем самым, «*X-у тошно*», «*X-i er flökurt*» — ситуативный признак. Характеристики самих предикатов *тошно* и *flökurt* сходны, но исл. *flökurt*, как видно по (17a) и (17b), сочетается как с актуальным, так и с генерическим субъектом, в то время как для русского языка второй вариант исключен.

Обозначим предикаты ситуативного признака, образованные от неглагольных основ, символом LexP, и рассмотрим LexP, соотнесенные с прилагательными. Производящие основы, от которых образуются предикативы и атрибутивные слова, делятся в русском языке на три класса: 1) актантно-поляризованные основы, от которых образуются только выражения, соотнесенные с конкретным референтом, и не образуются выражения, характеризующие ситуацию в целом, ср. запрет на **мне злобно, *мне упитанно, *мне аморально, *мне улыбчиво* при правильном *X — злобный, упитанный, аморальный, улыбчивый*[8]; 2) ситуативно-поляризованные

[8] Более точно, основы класса I делятся на два подкласса. От основ подкласса Ia предикативная несогласуемая форма с финалью *-о* вообще не может быть постро-

основы, от которых в литературном языке образуются только предикативы, ср. *X-у стыдно, тошно, совестно,* ср. запрет на **стыдный, *тошный, *совестный*; 3) амбивалентные основы, от которых образуются предикаты обеих типов — ср. *веселый, грустный, муторный, X-у весело, печально, муторно*. Пометим основы индексами I, II, III для амбивалентных, актантно-поляризованных и ситуативно-поляризованных основ соответственно. Отдельно рассмотрим случай, где лексическое значение LexP и коррелятивного прилагательного не тождественны, и предикатив реализует значение производящей основы с модификацией модального или оценочного плана. Если трактовать «коррелятивность» как диахроническое соответствие, такие примеры обнаружатся во всех трех классах основ, но в каждом из них они, вопреки В. В. Виноградову и А. В. Исаченко, составляют абсолютное меньшинство: утверждение, будто в русском языке образование предикативов от именных основ «всегда» или «как правило» сопряжено с изменением лексического значения, неверно. Результаты представлены в таблице на рис. 2.

на, ср. запрет на **это гневно, *это улыбчиво*. От основ подкласса Ib такая форма может быть построена, ср. *здесь пыльно, ?с его стороны любознательно проверить ссылки*, но валентность на субъекта-носителя признака, если таковая имеется, замещается не дат. п.лица, а предложным род. п. с предлогом *для*: *переписывать уже изданную книгу — для него (*ему) характерно*. Для целей данной статьи различение случаев Ia и Ib не важно.

Модификация лексического значения	I. Амбивалентные основы	II. Актантно-поляризованные основы	III. Ситуативно-поляризованные основы
	Образуются как предикаты, соотнесенные с конкретным референтом, так и предикаты ситуативного признака(LexP)	Образуются только предикаты, соотнесенные с конкретным референтом	Образуются только LexP
1. Нет	1.1. Скверн=ый → скверн=о $Adj_{-\varnothing, -а, -о, -ы}$ = {скверен, скверна, скверно, скверны}	2.1. гневн=ый ⊄ *мне гневн=о	3.1.*стыдн=ый → стыдн=о
2. Да	1.2. паршив=ый → ему сейчас паршиво	2.2. жирн=ый творог→ ему жирно	3.2. сла́б=ый → слаб=о́ ему

Рис. 2. Три типа соотношений краткое прил. → LexP в русском языке

В парах из последней строки таблицы (1.2., 2.2., 3.2.) лексическое значение производящей основы меняется. Предикат *X-у (сейчас) паршиво* не несет информации о множестве референтов, обладающих свойством *паршивости* в тот отрезок времени, когда физическое или психическое состояние X-а оценивается как *паршивое*. В этой паре негативная оценка *паршивых X-в /паршивой ситуации* как аномалии, крайней степени неблагоприятности, присутствует у обеих членов. Еще более очевиден сдвиг в паре *жирный творог → X-у жирно*: сообщение о том, что конкретные референты обладают свойством *жирности*, не имплицирует ни того, что жирность — негативная характеристика, ни того, что жирность — это градуируемый признак, ни того, что превышение порога жирности для каких-либо X-ов — аномалия. В литературном языке прилагательное *паршивый* и LexP *паршиво* в паре 1.1. одинаково приемлемы[9], что позволяет квалифицировать основу *паршив-* как амбивалентную. В паре 2.2. норма одобряет прилагательное *жирный* как обозначение свойств конкретных референтов, но не предикатив *X-у жирно* [*что-либо иметь/получить*] = *'По мнению Y-а, X не заслуживает Z-а'*. В консервативной норме основа *жирн-* актантно-поляризована. Тем не менее, большинство носителей нормы владеет и субнормативным употреблением LexP *жирно*. Это позволяет объяснить подтип 2.2. как спорадическое нарушение границ между амбивалентными (класс I) и актантно-поляризованными основами (класс II). То, что переход отдельных основ класса II в класс I — непродуктивный процесс, доказывается малым числом новых LexP, возникших за последние полвека. К подтипу 2.2. можно отнести пары *ленивый X → X-у лениво что-либо делать = 'По мнению Y-а, ленивая сущность X-а не позволяет ему делать Z'* и *фиолетовый X → X-у фиолетово* = 'X-у *все равно, Z или не Z*': в консервативной норме основы *ленив-* и *фиолетов-* актантно-поляризованы, а LexP *лениво* и *фиолетово* не разрешены. Если же ориентироваться на все регулярные употребления, то формы *жирно*, *лениво* и *фиолетово* необходимо признать. Точный список русских LexP зависит от соглашения о том, какого рода тексты разделяются большинством говорящих: не все, кто владеет книжным *X-у невместно*, употребляет разговорное *X-у лениво*, и наоборот. Существенно, что появление новых LexP связано с ротацией синонимов и вытеснением устаревших слов разговорными, а не с освоением новых лекси-

[9] Информанты, которые из пуристических соображений порицают предложение *Ему сейчас паршиво*, порицают и оценочное употребление прилагательного в предложениях типа *Это паршивый фильм*.

ческих значений, не покрывавшихся ранее существовавшими единицами: разг. *X-у лениво* пришло на смену нейтральному *X-у лень* и книжному *X-у невместно*, а жарг. *X-у фиолетово* является эпатажным вариантом разг. *X-у по барабану* и нейтральных *X-у все равно*, *X-у безразлично*.

Столь же непродуктивен подтип 3.2., ср. пару *сла́бый* → *X-у слабо́*, где сдвиг значения при образовании предиката ситуативного признака *X-у слабо́* = 'Y утверждает, что X не в силах выполнить задачу Z', сопровождается акцентной дифференциацией прилагательного и LexP. См. также пару *во́льный* → *X-у вольно́*, отмеченную А. М. Пешковским, и пару *по́лно <тебе притворяться>* и *полно́ <чего-л.>*. В этом подтипе этимологически связанные основы прилагательного и LexP можно трактовать как омонимы, что сближает его с подтипом 3.1., где производящая основа LexP исчезла: ранее существовавшие прилагательные **стыдный*, **совестный* больше больше не употребляются. Сторонники гипотезы о Категории Состояния как особой части речи — В. В. Виноградов, А. В. Исаченко, Н. С. Поспелов — уделили основное внимание непродуктивным механизмам, дающим подтипы 3.1 и 3.2., и изолятам, которые не связаны с прилагательными — *X-у надо*, *нельзя*, *жаль*, *пора*. Такой перекос продиктован желанием найти основы, для которых значение ситуативного признака является единственным. Но адъективных основ класса III в русском языке мало: для воспроизводства класса LexP важнее размежевание амбивалентных (класс I) и актантно-поляризованных (класс II) основ. Оно носит безусловный характер в литературной норме и относительно устойчиво в субнормативном русском языке.

Новейшие LexP с финалью *-о*, распространившиеся в молодежном жаргоне за последние десятилетия, уже не образуются от основ, дающих согласуемые краткие формы. Тем не менее, они коррелируют с парадигмами полных прилагательных так же, как LexP, уже вошедшие в словарь. Предикативы *X-у стремно, X-у очково, X-у ссыкотно* пока фиксируются не всеми интернет-словарями. LexP *стремно* связан с прилагательным *стремный* «опасный, рискованный». Основа *стремн-* у тех носителей русского языка, кто ей пользуется, амбивалентна и принадлежит к классу I. LexP *ссыкотно (сцыкотно)* 'X-у страшно до такой степени, что возникает позыв мочиться' связан с прилагательным *ссыкотный* (*сцыкотный*). Тем самым, и эта основа амбивалентна и принадлежит к классу I. LexP *очково* 'X-у страшно до такой степени, что возникает позыв непроизвольно испражняться' деривационно связан не с прилагательным *очковый*, образованным от существительного *очки* (ср. сочетание *очковая змея* в литературном языке), а с существительным *очко* 'anus' и устойчивой коллокацией *очко играет*. Тем самым, основу *очков-* следует отнести к классу III, если в субнормативном языке не всплывет прилагательное *очковый*$_2$, омонимичное слову литературного языка: в этом случае основу *очков*$_2$- следует отнести к классу I.

Функционирование LexP в субнормативном русском языке показывает, что появление новых предикативов, имеющих валентность на дат. п. лица, вполне вероятно, так ряд «X-у страшно до такой степени, что возникает физиологическая реакция Z_1, Z_2…Z_n» легко продолжить. В этом смысле класс LexP в русском языке открыт. Вместе с тем, количество предикативных значений, которые можно выразить синтаксической схемой ДПС [N_{dat} — V_{link} — Pred], ограничено денотативной сферой психических и физиологических состояний человека. Суммируем все обсужденные параметры:

Русский язык

- LexP, предицирующие таксономическое значение Состояния и соотнесенные с одушевленным семантическим субъектом, могут быть образованы только от одного класса качественных прилагательных.
- Семантика {+ Одушевленность; + эксперенциальный субъект; + актуальность; — генеричность}.
- Основная синтаксическая схема: ДПС [N_{dat} — V_{link} — Pred].
- Объем класса LexP: ок. 250–260 единиц, из 150–160 единиц образовы от адъективных основ.
- Продуктивность: класс LexP может пополнятьсся, но количество предикативных значений денотативно ограничено.

В древнеисландском языке предикативы со значением Состояния составляют открытый класс. Денотативное ограничение на одушевленность семантического субъекта отсутствует. Помимо дативных схем есть параллельные возможности реализации предикативов в предложении: значение «X-у не хватает хвороста» может быть выражено как путем *honum var / varð eldiviðarfátt*, букв. «ему было /стало малохворостно», с непереходными связками *vera* «быть» и *verða* «становиться»», так и путем *hann hafði eldiviðarfátt*, букв. «он имел малохворостно», с переходной связкой *hafa* "иметь" или ее синонимами (Циммерлинг 2002). В этих условиях разбиение производящих именных основ на актантно- и ситуативно-поляризованные принципиально невозможно, так как позицией семантического субъекта можно расширить любое высказывание с предикативом и связкой. Общее число форм типа *eldiviðarfátt*, которые имеют морфологию прилагательных, но употребляются только предикативно и лишены каких-либо согласовательных коррелятов (нет формы м. р.**eldiviðarfár* «малохворостный») в текстах

приближается к 1000, но некоторые встречаются в корпусе всего один-два раза. Как и в русском языке, синтактика и семантика предикативов, лишенных согласовательных коррелятов, ср. (18b) и (18c), и несогласуемых именных форм, ср. предикатив *illt* 'плохо' к прил. *illr* 'плохой' в (18a), идентична.

(18a)

þar var ***illt*** ***til eldiviðar*** (МН 15–198).
Там был$_{\text{3л. ед. ч. пр. вр.}}$ плохо к хворост$_{\text{род. п.}}$
«Там было ***плохо с хворостом***».

(18b)

því at matsveinum ***varð***
Так как повар$_{\text{дат. п. мн. ч.}}$ стать$_{\text{3л.ед. ч.}}$
eldiviðarfátt (М 15–200).
малохворостно$_{\text{пред.}}$
«Так как поварам не хватило хвороста», букв. «так как поварам ***стало малохворостно***».

(18c)

Matsveinar *Þorgeirs* ***höfðu***
"Повар$_{\text{им. п. мн. ч.}}$ Торгейр$_{\text{род. п.}}$ иметь$_{\text{3 л. Мн. ч.}}$
eldiviðarfátt (М 15–199).
малохворостно»
«У поваров Торгейра был недостаток хвороста», букв. «Повара Т. ***имели малохворостно***».

Морфология древнеисландских предикативов позволяет разбить их на три класса. Наиболее обширный класс — слова с финалью *-t* — строится по типу прилагательных т. н. сильного склонения в форме имен.–вин. п. ср. р. ед. ч. Среди них преобладают композиты и слова с суффиксами *-mikit*, *-fátt*, *-laust*, *-samt*[10]; также *-ult*, *-alt*[11], префиксами *ó-*, *ør-*, *auð-*, *vand-*, *tor-*[12]. Встречаются и симплексы: *annt* «много работы», букв. «занято», *ekki fritt* «не мирно» = «есть опасность вооруженного нападения». Второй класс представлен префиксальными словами, в качестве торого компонента которых выступает супин (несогласуемая форма причастия прош. врем.) от реально существующих глагольных основ: *auðfengit* «*легкодоступно» (*fá* «получать», **auðfá*), hugleikit «замышлено» (**hugleika*, *mér lék hugr* «у меня созрел замысел»), *ólífat* «не дожито» (*lifa* «жить», **ólifa*). По своему значению такие предикативы сопоставимы с целыми высказываниями, где значение глагола проясняется зависящими от него элементами: honum *kom* ***i hug*** «Х-у *пришло* ***на ум***» → *honum var* ***hugkvæmt,*** букв. «*Х-у было ***наумвходно***». Супин почти всех глаголов имеет финаль *-t,* тем самым большая часть предикативов получает общий маркер. Третий класс образуют слова с гласной финалью — *a/-i* (тип слабого прилагательного или *n*-основного существительного): *hugsi* «задумчив, молчалив», *þrotráða* «немощен», *ørbjarga* «беспомощен», *sundrorða* «несогласен, груб». Слова первых двух классов (с финалью-*t*) возможны в дативном предложении и его парафразах: тем самым, др. исл. предикативы на *-t* являются точной параллелью русских формам на *-о* (*стыдно, грустно* и т. п.) и прочим предикативам, формирующим русское ДПС (*жаль, надо, некогда* и т. п.), если отвлечься от того факта, что в русской грамматике нет парафраз, разрешающих преобразовывать непереходную схему *мне было стыдно/грустно*, в переходную **я имел стыдно/грустно*. Др. исл. предикативы на *-a/-i* обычно употребляются за пределами ДПС и реализуются в двусоставном предложении: они составляют точную параллель к русским предикативам типа *навеселе*, *подшофе*, *не в духе*.

Данные памятников показывают, что в древнеисландском языке XI–XIV существовало около 1000 ситуативно-поляризованных основ, от которых образовывались только предикативы ситуативного признака, но не обозначения постоянных свойств. Несогласуемые предикаты состояния с той же семантикой образовывались и от амбивалентных основ. Специфика древнеисландского языка заключается в том, что в нем, в отличие от современного русского, по-видимому, вообще не было актантно-поляризованных основ. Поскольку грамматика позволяет добавлять позицию субъекта-носителя признака в любую связочную структуру, нет явной грани между обозначениями состояний внешнего мира и обозначениями внутренних состояний лица: так, предикатив *vígljóst*, букв. «боесветло» сам по себе означает «достаточно светло, чтобы можно было вести бой», а при ДПС *honum var vígljóst*, букв. «ему$_{\text{дат. п. ед. ч.}}$ было$_{\text{3 л. ед. ч. пр. вр.}}$ боесветло» означает «Х-у было достаточно светло, чтобы он мог вести бой» и «В момент *t* было достаточно светло для того, чтобы Х мог вести бой». В этих условиях признак «± одушевленность» для класса предикативов (LexP) нерелевантен, тем более, что употребление предикативов в древнеисландском языке покрывает все денотативные сферы без ограничений. Суммируем параметры.

Древнеисландский язык

- LexP со значением Состояния образуются от любых основ. Актантно-поляризованных адъективных основ нет.
- Семантика {+ эксперенциальный субъект; + актуальность; — генеричность}. LexP не охарактеризованы как {+ одушевленные}.
- Основная синтаксическая схема: DPS [N_{dat} — V_{link} — Pred], связочные схемы с непереходными и переходными связками.

[10] Ср. др.исл. *áhyggjumikit*, букв. «озабоченно», *ráðfátt* «нет мыслей», букв. «*скудномысленно», *svefnsamt* «(делается) сонно».

[11] Др. исл. *stopalt* «скверно», др. исл. *bimbult* «муторно».

[12] Др. исл. *ørvænt* «безнадежно», др. исл. *torfynt* «трудно найти».

- Объем класса LexP: ок. 700–1000 специализированных в предикативном употрблении единиц в корпусе памятников.
- Продуктивность: LexPs — открытый класс, набор предикативных значений не ограничен.

В современном чешском языке предикативы со значением Состояния составляют строго ограниченный класс (не более 25–30 единиц), его пополнение невозможно. Предикативы, которые реализуются в ДПС, лишены единого маркера: наряду с формами ср. р. бывших кратких прилагательных (финаль *-о*), имеются предикативные наречия с финалями — é (*trapné* «грустно», *lhostejné* «безразлично») и -е (*dobre* «хорошо», *zle* «плохо»), и слова с финалью *-а* (*zima* «холодно», *hanba* «стыдно»). Другой особенностью чешского языка является то, что краткое прилагательная как живая категория отсутствует: сохранившиеся 30–35 согласуемых кратких прилагательных имеют статус архаизмов. Предикативы с финалью *-о* в диахронической перспективе являются формами имен-вин. п. ср. р. кратких прилагательных, но в синхронии коррелятивность утрачена. Те формы на *-о,* которые употребляются в ДПС, ср. чеш. *je mi smutno* «мне грустно», *je mi nevolno od žaludku* «мне муторно», *je mi slabo* «мне тошно, дурно», *je mi teskno* «мне тоскливо», je *mi* lehko «мне весело», не соотнесены с согласуемыми краткими прилагательными, и наоборот, от тех кратких прилагательных, которые еще фиксируются книжной нормой (*bos* «бос», *ženat* «женат», *zvědav* «любопытен», *bohat* «богат» и т. п.), не образуются предикативы, употребляемые в ДПС. Тем самым, чешские краткие прилагательные и чешские предикативы *-о*, образованные от кратких прилагательных, образуют дополнительную дистрибуцию: первые представляют актантно-поляризованными основы (ср. рус. *смешлив-*), а вторые — ситуативно-поляризованные (ср. рус. *стыдн-*). Амбивалентных основ типа рус. *смешн-*, *грустн-*, обеспечивающих продуктивность ДПС в русском, в чешском языке почти нет.

Еще одна особенность чешского языка состоит в том, что не все несогласуемые предикативы, которые употребляются в ДПС (схема N_{dat} — V_{link} — Pred, ср. рус. ***мне*** *грустно, что P*) могут реализоваться в схеме с нулевым подлежащим (∅— V_{link} — Pred, ср. рус. ∅ *грустно, что P*). Так, рус. *жаль* в чешском языке соответствует два разных предикатива. Значение «X-у жаль Y-а» передается в ДПС с помощью *líto — je* ***mi*** *líto matky /peněz* «***мне*** жаль мать/денег», а значение «X-у жаль, что *P*» передается с помощью *škoda — je škoda, že P.* Употребления **je líto matky* и **je mi škoda, že P* неграмматичны. Точно также, при *teskno* «тоскливо», *těžko* «тяжко», *nevolno* «муторно», *slabo* «дурно», *smutno* «грустно», *trapné* «неловко» позиция дат. п. лица неустранима: эти предикативы не сочетаются ни с формальным словом *to* “это”, ни с нулевым подлежащим: передать смысл «*это* было грустно», «*это* было тоскливо» в схеме ∅/*to* — V_{link} — Pred — **∅/*to bylo teskno*, **∅/*to bylo smutno* — нельзя. И наоборот, предикатив *veselo* «весело», который разрешает нулевое подлежащее, ср. ***tady*** *je veselo* «***здесь*** весело», несовместим с ДПС, предложение **je* ***mi*** *veselo* «мне весело» отвергается. Предикативы *lze* «разрешено», *není lze* «нельзя», *možno* «можно» не принимают дат. п. лица: используя эти модальные операторы, по чешски нельзя сказать «мне нельзя», «мне можно». Таким образом, в чешском языке за ДПС и схемой с нулевым подлежащим закреплены разные группировки предикативов, что русскому и древнеисландскому языку несвойственно.

Некоторые предикативы допускают или требуют транзитивную связку *mít* «иметь», но такие случаи лексикализованы. Значение «мне нужно» при предикативе *třeba* в высоком стиле можно выразить в ДПС — чеш. *je* ***mi*** *to třeba čeho*, чему в стандартном языке соответствует транзитивный оборот с синонимом *potřeba* «нужно, потребно» — чеш. ***mám*** *potřebu čeho*, букв. «***я имею*** нужно в чем-л.», но *potřeba* не имеет валентности на дат. п. лица — **je* ***mi*** *to potřeba čeho,* а *třeba* не сочетается с *mít* — **mám potřebu*. Только с *mít* реализуется предикатив *rád* «рад», «X расположен к Y-у»[13]: Предикатив *dusno* «душно» реализуется и в ДПС, и с mít, но ***mám*** *dusno*, букв. «***я имею*** душно», значит не «мне душно», а «здесь душно», информация о состоянии одушевленного субъекта — «X-у душно» в схеме с *mít* передана быть не может. Суммируем параметры.

Современный чешский язык

- Краткие прил. отмирают. Лишь часть LexP имеет соответствия в парадигмах кратких прилагательных. LexP образуются от малого числа адъективных основ. Амбивалентные основы отсутствуют.
- Семантика {+ эксперенциальный субъект; + актуальный; — генерический}. LexP не охарактеризованы как {+ одушевленные}.
- Основные синтаксические схемы: ДПС [N_{dat} — V_{link} — Pred], ПС [∅ — V_{link} — Pred]. Разные схемы закреплены за разными группировками предикативов.
- Объем класса LexP: ок. 25–30 единиц.
- Продуктивность: LexP — закрытый класс, количество предикатных значений строго ограничено.

[13] Предикатив *potřeba* изменяется по падежам (точнее, принимает формы им. п. и вин. п. ед. ч. при непереходной и переходной связках соответственно), предикатив *rád* имеет формы рода и числа (но не падежа). Ни то, ни другое слово не могут замещать актантные позиции и принимать определения, т. е. с синтаксической точки зрения не являются существительными.

Подводя итоги, можно сделать вывод, что между наличием связочных дативно-безличных структур и наличием предикативов со значением состояния (LexP) есть связь, но общность синтаксической схемы позволяет предсказать только общую семантику, но не их число и синтактику LexP. В современном русском языке между двумя основными продуктивными моделями дативных предложений — Дативно-Предикативными Структурами (ДПС) и Дативно-Инфинитивными Структурами (ДИС) сложилось такое распределение, которое позволяет утверждать, что они нигде не являются синтаксическими синонимами. В русском и древнеисландском языках предикативы получают в ДПС общие маркеры, указывающие на их генетическую отнесенность к парадигмам прилагательных. Основы, от которых образуются предикативы, могут быть либо специализированы для построения ситуативных обозначений, либо амбивалентны и использоваться как для производства слов ситуативного признака, так и для обозначения постоянных свойств, имеющих конкретного носителя. В русском языке четко прослеживается тернарное противоставление: ситуативно-поляризованных адъективных основ класса *стыдн-* мало, имеется большое число актантно-поляризованных основ, от которых предикативы ситуативного признака образовываться не могут, а продуктивность ДПС поддерживается амбивалентными основами класса *грустн-*. В древнеисландском языке нет актантно-поляризованных основ, в то время как в чешском языке нет амбивалентных основ. Что касается лексических значений, выражаемых предикативами, были выявлены как языки, где в ДПС выражается строго ограниченное число значений (чешский), так и языки, где в ДПС выражается неограниченно большое число значений (древнеисландский). Семантическим инвариантом предикативов является значение локализованного во времени признака ситуации: в зависимости от конкретного языка, данный признак может либо соотноситься, либо не соотноситься исключительно с одушевленным субъектом. Во втором случае дативные структуры с позицией именного предикатива обнаруживают точки сближения с модальными глагольными конструкциями (ср. латышский язык).

Литература

1. *Апресян Ю. Д.* Синтаксические признаки лексем // Russian Linguistics, Vol. 9, N 2–3, 1985, 280–315.
2. *Бонч-Осмоловская А.* Конструкции с дативным субъектом в русском языке. Диссертация на соискание ученой степени кандидата фил.наук. М., МГУ, 2003.
3. *Булыгина Т. В.* К построению типологии предикатов в русском языке // Семантические типы предикатов, М., 1982, 7–85.
4. *Виноградов В. В.* Русский язык. Грамматическое учение о слове. М.-Л., 1947.
5. *Виноградов В. В.* Из истории изучения русского синтаксиса. М., МГУ. 1958.
6. *Исаченко А. В.* О возникновении и развитии "категории состояния" в славянских языках // Вопросы языкознания, 1955, N 6.
7. *Всеволодова М. В.* Теория функционально-коммуникативного синтаксиса. Учебник. М., МГУ: 2000.
8. *Грамматика* 1980 — Русская грамматика. АН СССР, т. 1–2. М.: Наука, 1982.
9. *Гришина Н. И.* Дативные предложения в парадигматическом аспекте. М., 2002.
10. *Золотова Г. А.* Коммуникативные аспекты русского синтаксиса. М., 1982.
11. *Исаченко А. В.* О возникновении и развитии "категории состояния" в славянских языках // Вопросы языкознания, 1955, N 6.
12. *Мельчук И. А.* Syntactic, or Lexical Zero in Natural Language // Модель Смысл-Текст и проблемы русской грамматики. Wiener Slawistischer Almanach. Sonderband 39. Москва-Вена, 1995, 169–205.
13. *Поспелов Н. С.* В защиту категории состояния // Вопросы языкознания, 1955, № 2.
14. *Селиверстова О. Н.* Второй вариант классификационной сетки и описание некоторых предикатных типов русского языка // Семантические типы предикатов, М., 1982, 86–157.
15. *Щерба Л. В.* Языковая система и речевая деятельность. М.- Л., 1974.
16. *Циммерлинг А. В.* История одной полемики // Язык и речевая деятельность, 1998, № 1, 63–88.
17. *Циммерлинг .В.* Древнеисландские предикативы и гипотеза о категории состояния // Вопросы Языкознания, 1998, № 2.
18. *Циммерлинг А. В.* Субъект состояния и субъект оценки // Логический анализ языка. Образ человека в зеркале языков и культур/Арутюнова Н. Д. (ред.). М., Индрик, 1999.
19. *Циммерлинг А. В.* Типологический синтаксис скандинавских языков. М.: Языки славянской культуры, 2002.
20. *Циммерлинг А. В.* Два типа дативных предложений в русском языке // Слово — чистое веселье: Сборник статей в честь А. Б. Пеньковского М.: Языки славянских культур, 2009, 471–489.
21. *Andronov A. V.* “Vajazdības izteiksme” latviešu valodas gramatiskajā tradīcijā (The fluctuating fortunes of the Latvian debitive) // Baltu filoloģija 8 (1999), 154–177.
22. *Hoelvoet A.* Studies in the Latvian Verb. Kraków? 2001.
23. *Moore, J. and D. Perlmutter.* What Does it Take to Be a Dative Subject? // Natural Language and Linguistic Theory 18 (2000), 373–416.
24. *Sigurðsson, Halldór Á.* To be an Oblique Subject: Russian vs. Icelandic // Natural Language and Linguistic Theory 20 (2002), 691–724.
25. *Zatovkaňuk M.* Neosobní predikativa a utváry přibuzné, zvlaště v ruštine //Rozpravy Československé akademie věd, 1965, sešit 6, ročník 75.
26. *Zimmerling A.* Die unpers`nlichen Satzmodelle in der altisl@ndischen Sprache // Texte und Untersuchungen zur Germanistik und Skandinavistik, Bd., 30, Lilja Popowa (hrsg). Peter Lang Verlag, 1992. Frankfurt a.M.- Berlin- Bern — Paris- New York, 309–328.
27. *Zimmerling A.* Zero Lexemes and Derived Sentence Patterns // Wiener Slawistischer Almanach, Sonderband 69 (2007).
28. *Zimmerling A.* Dative Subjects and Semi-Expletive pronouns // Studies in Formal Slavic Phonology, Syntax, Semantics and Information Structure / G. Zybatow, Uwe Junghanns, Denisa Lenertová, Petr Biskup (eds). Frankfurt-a-M-Berlin-Bern-Bruxelles-N.Y –Oxford-Wien: Peter Lang, 2009, 253–265.

Об одном статистическом методе пополнения морфологического словаря

Yet another statistical method for non-vocabulary word flexion prediction based on text corpora

Черненьков Д. М. (drcheren@gmail.com)

Московский институт электроники и математики, Москва, Россия.

В статье предлагается алгоритм предсказания парадигм словоизменения несловарной лексики из текстовых корпусов. В основе алгоритма лежит ряд вероятностных моделей словоизменения, а также метод машиннообучаемого отбора и ранжирования объектов. В статье проведен анализ статистических свойств корпуса и результатов обучения модели.

Введение

Создание и пополнение морфологического словаря для флективных языков крайне трудоемко по причине сложности правил русской морфологии. Однако существует возможность полностью или частично автоматизировать этот процесс средствами анализа текстов, содержащих несловарную лексику.

Существует несколько методов создания автоматических морфологий. Требования к морфологическому анализатору зависят от задачи, для решения которой он применяется. Так, в [1] описаны методы морфологического анализа для поисковых систем. Основными требованиями в данном случае являются производительность и полнота анализа. Т. е. анализатор должен выдавать результат для максимального количества различных словоформ. При этом подробная информация, как правило, не требуется. Для основных задач информационного поиска достаточно получения нормальной формы из произвольной словоформы. В такого рода анализаторах словарь не является обязательным компонентом, и даже при его наличии производится анализ как словарных, так и несловарных словоформ. В [2,3] описаны варианты алгоритмов морфологического анализа, использующие индексы постфиксов, и набор правил словообразования.

Системы, в которых морфологический анализ является лишь промежуточной ступенью, за которой, как минимум, следует ступень синтаксического анализа [4,5], нуждаются в более подробной грамматической информации о каждой словоформе. В русском языке можно насчитать до 15 грамматических категорий, так или иначе выразимых морфологически, до 100 грамматических значений на часть речи (и соответственно, до 100 форм на слово) и свыше 2000 парадигм словоизменения [6]. В связи с этим, для полного, точного и подробного морфологического анализа русских словоформ необходим словарный алгоритм. Причем словарь должен содержать полную грамматическую информацию о каждом слове и каждой словоформе.

Большинство словарных морфологий для русского языка основано на грамматическом словаре А. А. Зализняка. Однако для применения анализаторов в практических целях необходимо пополнение словаря тем или иным методом. В [7,8] описаны методы автоматического и автоматизированного пополнения морфологических словарей с использованием корпусов текстов. Эти методы основаны на статистических свойствах языка. Основной проблемой, которая решаются в этих работах, является проблема неоднозначности разбора несловарных словоформ и необходимости выбора правильной гипотезы лемматизации. В [7] предлагается группировать гипотезы различными методами и выбирать группы с наибольшей встречаемостью словоформ в корпусе. В [8] данный метод дополняется учетом статистических свойств парадигм, а также учет наиболее используемых словообразовательных префиксов.

Оба метода дают достаточно высокую точность. Однако в [7] выходным результатом алгоритма является набор гипотез, из которых по-прежнему необходимо выбирать верную, т. е. процесс пополне-

ния словаря автоматизирован лишь частично. В [8], с другой стороны, на выходе мы получаем лишь канонические формы для несловарных словоформ, в то время как для многих задач необходимо определить полную парадигму словоизменения.

Кроме того, при решении задачи снятия омонимии (при которой возникает схожая проблема неоднозначности морфологического анализа) используются методы анализа ближайшего контекста анализируемой словоформы [9]. В основе метода лежит предположение, что соседние слова с большой вероятностью связаны грамматическими отношениями, такими как согласование или управление. Данный метод можно применять и для выбора гипотез лемматизации несловарных словоформ.

1. Цели и задачи

Целью данной работы является объединение существующих методов анализа несловарной лексики и пополнения словарей для создания максимально точного алгоритма пополнения, требующего минимальной ручной обработки результатов. За основу берется математическая модель и алгоритм, описанные в [10]. Кратко этот алгоритм можно описать как последовательность шагов:

1. Выделить несловарные словоформы из корпуса
2. По каждой словоформе построить все возможные гипотезы лемматизации. Объединить построенные гипотезы в одно множество без дубликатов
3. Отфильтровать гипотезы по некоторому признаку
4. Кластеризовать гипотезы, выделив компоненты связности в биграфе гипотезы-словоформы
5. Из каждого класса по некоторому критерию выбрать одну или несколько наилучших гипотез

Для достижения поставленной цели необходимо решить ряд проблем:

1. Оценка качества и анализ результатов. Отсутствие достаточно крупного размеченного корпуса с несловарной лексикой. Кроме того, пока неясен критерий оценки точности анализа. В приведенных во введении работах используются различные критерии качества. Необходима выработка критерия качества в соответствии с основным назначением данной системы.
2. Большое число признаков, которые нужно учитывать при отборе гипотез. Затруднителен ручной подбор критериев отброса и ранжирования гипотез.
3. Большое количество парадигм в словаре порождает множество гипотез для каждой словоформы. Среди этих гипотез лишь небольшая доля верных (около одной тысячной). При попытке анализа признаков совокупность верных гипотез оказывается статистически незначимой.

Проблема отсутствия размеченного корпуса решается разделением множества лексем существующего словаря на генерирующее и валидационное подмножества (с образованием соответственно двух словарей, генерирующего и валидационного, с общим набором парадигм). При этом имитируется ситуация пополнения словаря: множество анализируемых словоформ включает в себя все словоформы, которые не входят в генерирующий словарь, но входят в валидационный. Далее, во всех местах алгоритма, где обычно используется полный словарь, теперь используется генерирующий. Назначение валидационного словаря — проверка сгенерированных гипотез (в силу описанного алгоритма выбора словоформ, правильный их разбор всегда можно определить из словаря).

2. Сокращение числа гипотез

В связи с большим числом парадигм в словаре (около 2700), на каждую несловарную словоформу генерируется несколько тысяч гипотез. Такое большое количество гипотез, во-первых, серьезно сказывается на производительности, а во-вторых, делает долю правильных гипотез очень низкой, что затрудняет процесс машинного обучения. Для снижения числа неверных гипотез предпринимаются следующие меры:

- Кластеризация парадигм. Если у парадигм A и B одинаковые части речи и наборы неизменяемых параметров, и при этом множество форм парадигмы A целиком входит во множество форм B, то эти две парадигмы объединяются, а при генерации гипотез используется только B (с большим числом форм). Этот метод позволяет сократить число парадигм, а значит и генерируемых гипотез, приблизительно вдвое. Поскольку гипотезы считаются эквивалентными, верные разборы не теряются.
- Фильтрация гипотез перед кластеризацией. Подробнее описана в [10].
- Отсечение гипотез в кластерах. Для каждого кластера составляется список всех словоформ, покрываемых его гипотезами. Из этих словоформ выбирается 3, наиболее часто входящие в корпус. Гипотезы, не покрывающие хотя бы одну из этих словоформ, удаляются из кластера. При этом отсеивается около 60 % неверных гипотез и около 5 % верных. Данная эвристика предложена в [13], где она используется для сокращения размера поискового индекса.

Даже с учетом этих мер число неверных гипотез в выборке превышает число верных более, чем в 100 раз. Построение качественного регрессора на такой выборке затруднительно.

3. Применение машинного обучения

Задачи фильтрации и выбора объектов некоторого типа по множеству признаков можно сформулировать как задачи машинного обучения при условии наличия подходящей обучающей выборки. Кроме того, необходимо количественное выражение всех признаков отбора. Если использовать метод деления словарей, описанный выше, можно построить достаточно большую выборку гипотез, помеченных как либо верные, либо неверные. При этом для фильтрации можно использовать алгоритмы классификации (разделение на 2 класса — верные и неверные гипотезы), а для выбора гипотез из кластеров — алгоритмы регрессии (генерация метки-действительного числа для каждой гипотезы и выбор из каждого класса гипотез с наибольшими значениями меток).

Однако при построении классификатора для фильтрации возникает следующая проблема: в обучающей выборке присутствует хотя бы одна верная гипотеза для каждой анализируемой словоформы, в то время как основное назначение фильтрации — отбросить опечатки и прочий «мусор», т. е. токены, в принципе не имеющие верного разбора. Поэтому на данный момент предлагается использовать для фильтрации старые критерии, полученные в [10].

Проблема выбора верной гипотезы из кластера не сводится непосредственно к задачам классификации, регрессии и кластеризации, которые решаются основными методами машинного обучения. Попытка классифицировать гипотезы на верные и неверные приводит к очень низкой точности, поскольку мощность класса неверных гипотез приблизительно на 2 порядка больше мощности класса верных гипотез. В результате точность обычных методов классификации в применении к этой задаче не превышает 3–5 %, что лишь немного выше случайных результатов.

Скорее, данная проблема выбора гипотез аналогична проблеме ранжирования, которая наиболее распространена в системах поиска информации. Отличие лишь в том, что задача ранжирования состоит в сортировке элементов в правильном порядке, а описываемая задача — в выборе наилучшего из них. В [14] описаны два подхода к ранжированию:

1. Парное ранжирование. Задача сравнения ранжируемых объектов сводится к задаче классификации. Элементами выборки для обучения классификатора являются пары ранжируемых объектов, пары разделяются на 2 класса: те пары, в которых первый объект лучше второго, и те, в которых второй объект лучше первого. Поиск наилучшего объекта тоже может легко осуществляться этим методом. Однако данный метод обладает рядом недостатков. Во-первых, далеко не все алгоритмы классификации гарантируют необходимые свойства функции сравнения (транзитивность и антисимметричность). Во-вторых, для получения удовлетворительной точности выбора необходима очень высокая точность классификатора. Так, для выбора верного элемента из 10 со средней точностью 50 % необходима функция сравнения, работающая с точностью не менее 97 %. С ростом числа ранжируемых элементов допустимый процент ошибки классификатора падает с экспоненциальной скоростью.
2. Списковое ранжирование. Выбирается оценочная функция, ставящая в соответствие каждому объекту действительное число — оценку. При ранжировании объекты сортируются в порядке убывания оценок. При выборе просто берется объект с наивысшей оценкой. Этот подход и использован в данной работе.

Предлагаемый метод выбора гипотез из кластера основан на алгоритме ранжирования, представленном в [15]. Выбор наилучшей гипотезы из списка производится функцией y = h(X),где X — конечное множество гипотез $X=\{x_1, x_2, \ldots, x_n\}$, y –гипотеза, $y \in X$.

Для оценки гипотез вводится скалярная функция f(x), где x — гипотеза. Конкретный вид функции в данном разделе не важен. Тогда функция h имеет вид:

$$h(\{x_1, x_2, \ldots x_n\}, f) = \operatorname*{argmax}_j f(x_j), \quad j = \overline{1,n}$$

Задача состоит в нахождении функции f*, обеспечивающей максимальную точность выбора гипотез, т. е. если в произвольном кластере гипотез X верной является y, то

$$f^* = \operatorname*{argmax}_f P(h(X, f) = y$$

Естественно, в распоряжении имеется лишь ограниченная случайная выборка пар $S=\{\{X_1,y_1\},\ldots \{X_m,y_m\}\}$, с помощью которой необходимо оценивать точность выбора и оптимизировать функцию f. Для этого вводится приблизительная оценка точности L(S, f). Тогда мы можем определить ожидаемую функцию $\hat{f} = \operatorname{argmax}_f L(S, f)$. По аналогии с предложенной в [15] оценкой логарифмического правдоподобия для ранжирования, предложим функцию для оценки точности выбора:

$$L(S,f) = \sum_{i=1}^{m} \log\left(\hat{P}(h(X_i) = y_i)\right),$$

где

$$\hat{P}(h(X_i) = y_i) = \frac{\exp\left(f(y_j)\right)}{\sum_{j=1}^{|X_i|} exp(f(-x_i^j))}.$$

Эта оценка аналогична оценке правдоподобия, используемой в логистической регрессии. Согласно [16], такая оценка дает хорошее приближение, если распределение величины f относится к семейству экспоненциальных распределений. Наиболее распространенные в статистических задачах распределения, такие как Нормальное и биномиальное распределения, распределения Пуассона и Бернулли. Кроме того, функция L дифференцируема, что позволяет использовать градиентные методы для нахождения f.

4. Описание и анализ признаков

Для представления гипотез в функции оценки необходим набор количественных характеристик гипотезы. В [10] был предложен ряд признаков: частотность словоформ, покрываемых гипотезой, в корпусе, число различных покрываемых словоформ в корпусе, число лексем в словаре с тем же постфиксом псевдоосновы заданной длины. В данной работе предлагается модифицировать и расширить набор признаков. Ниже приводится полный список с описаниями:

1. Число вхождений в корпус словоформ, покрываемых данной гипотезой. В качестве
2. Число различных словоформ входящих в корпус и покрываемых данной гипотезой
3. Число лексем в словаре с той же парадигмой, и имеющих тот же постфикс длины l, что и словоформы данной гипотезы, причем берется среднее арифметическое этих чисел для всех форм. Разным значениям l соответствуют разные признаки
4. Оценка вероятности вхождения грамматических форм в корпус. Пусть по гипотезе можно построить словоформы $w_1, w_2, \ldots w_n$. Им соответствуют грамматические значения $p_1, p_2, \ldots p_n$. (Грамматическое значение определяется частью речи и полным набором параметров лексемы и словоформы.) Значение данного признака считается по формуле $\sum_i F(w_i)\log\left(F(p_i)\right)$.
5. Оценка вероятности вхождения биграмм грамматических форм в корпус. Для каждого вхождения словоформ вычисляется частота вхождения в корпус биграмма грамматического значения этой словоформы и предшествующей ей в тексте словоформы. В качестве значения признака используется сумма логарифмов этих частот. Аналогичный признак вычисляется с учетом последующей словоформы, а не предыдущей.

Для исследования признаков был использован корпус новости с портала rbc.ru за период с января 2003 по декабрь 2008. Распределения значений признаков показаны на диаграммах приведенных ниже.

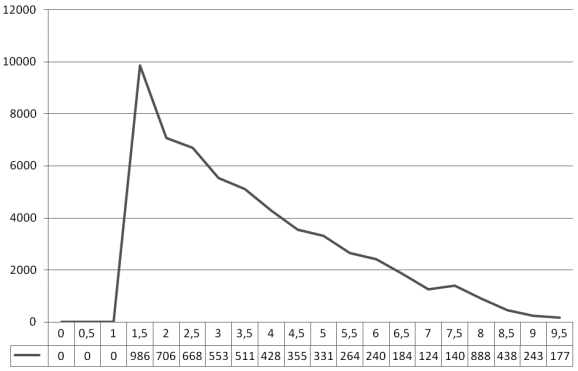


Рис. 1. Распределение частотности словоформ гипотезы

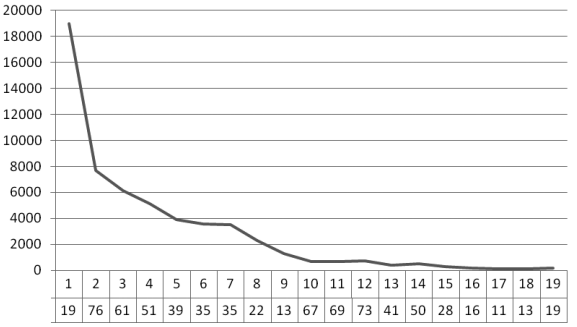


Рис. 2. Распределение числа встреченных словоформ гипотезы

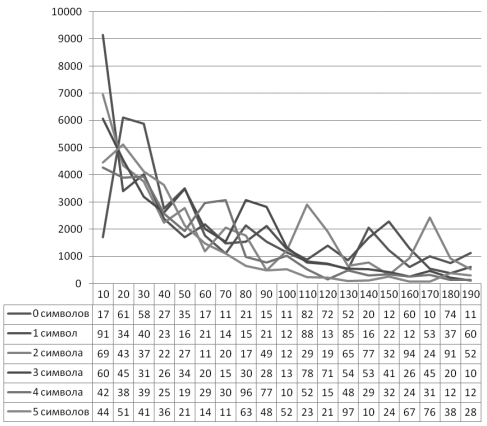
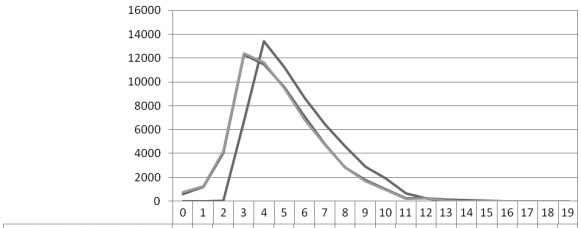


Рис. 3. Число лексем парадигмы с тем же постфиксом заданной длины



	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Отдельные словоформы	0	0	68	67	13	11	86	64	46	29	19	65	25	14	86	37	1	0	0	0
Биграммы по префиксам	63	12	40	12	11	95	70	47	28	17	99	22	24	9	0	0	0	0	0	0
Биграммы по суффиксам	75	12	42	12	11	94	68	47	28	16	94	18	22	42	0	0	0	0	0	0

Рис. 4. Распределение оценок встречаемости отдельных словоформ и биграммов

5. Обучение и тестирование классификатора

В качестве функции f для ранжирования гипотез использовалась линейная комбинация признаков с коэффициентами θ. Для оптимизации значения коэффициентов был использован стохастический градиентный спуск [11]. Градиент рассчитывался как $\nabla_\theta L(S, f)$. Точность отбора в таком виде оказалось очень низким — 13 %.

Тогда был применен алгоритм пошагового отбора признаков. В результате качество повысилось до 37 %, а среди признаков осталось только три: число различных словоформ и вероятности биграмм (как с предшествующим, так и последующим словом).

Кроме того, была оценена вероятность нахождения правильных гипотез, если брать несколько гипотез:

Число отобранных гипотез	Вероятность найти верную гипотезу
1	37 %
2	49 %
3	55 %
4	59 %
5	60 %
6	60 %
7	61 %
8	62 %
9	62 %
10	62 %

Выводы

Разработан новый алгоритм предсказания модели словоизменения несловарной лексики из текстовых корпусов. Точность полностью автоматического отбора составляет 37 %.

Несмотря на то, что низкая точность не позволяет использовать данный алгоритм для полностью автоматического пополнения словаря, он позволяет существенно облегчить ручной метод заполнения. Аналогичных опубликованных результатов по заполнению словаря с точным соответствием полной парадигмы словоизменения найдено не было.

Отбор признаков показал, что наиболее важными критериями гипотез о словоизменении являются результаты анализа грамматического окружения вхождений несловарных словоформ. В дальнейших исследованиях необходимо провести анализ более дальнего окружения, чем соседние слова.

Литература

1. *Сегалович И. В., Маслов М. А.* Русский морфологический анализ и синтез с генерацией моделей словоизменения для не описанных в словаре слов // М.: Диалог, 1998.
2. *Гельбух А. Ф.* Эффективно реализуемая модель морфологии флективного естественного языка // М.: Всероссийский институт научной и технической информации, 1994.
3. *Ножом И. М.* Морфологическая и синтаксическая обработка текста (модели и программы), // М.: 2003,
4. *Крылов С. А., Старостин С. А.* Актуальные задачи морфологического анализа и синтеза в интегрированной информационной среде Starling // М.: Диалог, 2003.
5. *Мальковский М. Г., Старостин А. С.* Система морфосинтаксического анализа Treeton и мультиагентный синтаксический анализатор Treevial: принцип работы, система правил и штрафов // Екатеринбург: Изд-во Уральского университета, 2007. — С. 135–143.
6. *Зализняк А. А.* Грамматический словарь русского языка, 2-е изд. // «Русский язык», М.: 1980.
7. *Ляшевская О. Н., Сичинава Д. В., Кобрицов Б. П.* // Екатеринбург: Изд-во Уральского университета, 2007. — С. 118–125.
8. *Андреев А. В., Березкин Д. В., Симаков К. В.* Обучение морфологического анализатора на большой электронной коллекции текстовых документов. Труды седьмой всероссийской научной конференции — Ярославль: Ярославский государственный университет, 2005. — С. 173–181.
9. *Сокирко А. В.* Сравнение эффективности двух методик снятия лексической и морфологической неоднозначности для русского языка // М.: 2005
10. *Черненьков Д. М.* Предсказание морфологических характеристик и парадигм словоизменения несловарных словоформ в текстах на русском языке // Киев: Труды конференции Мегалинг, 2009
11. *Alpaydin E.* Introduction to Machine Learning // The MIT Press 2004
12. *Friedman J., Hastie T., Tibshirani R.* The Elements of Statistical Learning: Data Mining, Inference, and Prediction // Springer, Stanford 2009
13. *Очагова Л. Н., Маслякова В. М., Зайцева Е. М., Дунаевская С. М.* Универсальная технология формирования словаря баз данных CDS/ISIS с использованием основ терминов // Государственная публичная научно-техническая библиотека России, М.: 2000
14. *Cao Z., Qin T., Liu T. Y., Tsai M. F., Li H.* Learning to rank: From pairwise approach to list wise approach // Proceedings of the 24th International Conference on Machine Learning. Corvallis, OR, 2007
15. *Xia F., Liu T. Y., Zhang J. W., Li H.* Listwise Approach to Learning to Rank — Theory and Algorithm // Proceedings of the 25th International Conference on Machine Learning. Corvallis, OR, 2008
16. *Banerjee A.* An Analysis of Logistic Models: Exponential Family Connections and Online Performance // Proceedings of the Ninth Workshop on Algorithm Engineering and Experiments (ALENEX), 2007

Автоматическое извлечение оценочных слов для конкретной предметной области

Automatic extraction of domain-specific opinion words

Четвёркин И. И. (ilia2010@yandex.ru)

Факультет вычислительной математики и кибернетики МГУ

Лукашевич Н. В. (louk_nat@mail.ru)

НИВЦ МГУ

Для эффективного извлечения мнений из текстов необходимо знание оценочных слов и выражений из рассматриваемой предметной области. Мы предлагаем новый подход к автоматическому извлечению оценочных слов, основанный на работе с несколькими корпусами текстов и вычислении с их помощью характеристик слов.

1. Введение

В настоящее время на страницах сети Интернет можно найти множество отзывов о тех или иных товарах, фильмах, книгах и т. п. Эти отзывы содержат много полезной информации, поэтому ее важно автоматически извлекать и предъявлять пользователям [6, 11].

Мнения пользователей о продукте часто выражаются посредством оценочных слов и выражений, которые несут в себе некоторую положительную или отрицательную оценку. Поэтому важным фактором качественного извлечения мнений о той или иной сущности является знание оценочных слов и выражений, которые используются в данной области. Проблема состоит в том, что невозможно заранее собрать список оценочных слов и выражений, которые будут применимы для всех предметных областей, поскольку некоторые оценочные выражения употребляются только в конкретных предметных областях, другие являются оценочными в одной области и не являются оценочными в другой.

Так, например, выражение «хочу еще сходить посмотреть» является характерным оценочным выражением для фильмов и небольшого количества других областей. Значимую часть оценочных слов не найти ни в каких словарях, например, «никакущий», Иногда трудно догадаться, что то или иное слово в контексте может употребляться как оценочное, как, например, слово «скомканный» о фильмах или книгах.

Таким образом, актуальной является задача автоматического формирования списка наиболее употребительных оценочных слов для данной предметной области.

В данной статье мы рассмотрим методы автоматического получения оценочных слов на основе нескольких корпусов текстов, которые можно автоматически построить для многих предметных областей, а именно, корпуса отзывов о сущности с вручную проставленными потребителями оценками, корпуса нейтральных описаний сущностей и нейтрального контрастного корпуса, например, составленного из потока общезначимых новостей.

Из указанных корпусов мы извлекаем списки слов, упорядоченные по значению различных признаков, оцениваем качество этих списков относительно содержания в них оценочных слов, и исследуем способы комбинирования этих списков для получения лучшего качественного состава по оценочным словам. Определение ориентации оценочных слов не производится.

Рассматривается предметная область отзывов о фильмах.

2. Методы автоматического извлечения оценочных слов

Существует два основных подхода к автоматическому выделению оценочных слов из текстов. Первый

подход базируется на информации из словарей или тезаурусов. В данном подходе обычно выбирается небольшое начальное множество слов, которое формируется вручную, и затем дополняется с помощью словарей и тезаурусов. Основной принцип заключается в том, что если слово оценочное, то и его синонимы будут оценочными и антонимы (возможна только смена ориентации), Поэтому, имея слова из начального множества, можно с помощью этих связей составить новое множество, которое будет более полным [5], В [3] на основе толкований слов в словаре выясняется их ориентация (положительная или отрицательная), Основная идея заключается в том, что слова с одинаковой ориентацией имеют «похожие» толкования. Таким образом, основываясь на этой идее, был построен классификатор слов на положительно ориентированные слова и отрицательно ориентированные.

Корпусный подход основан на поиске правил и закономерностей в текстах. В работе [8] оценочная характеристика слова вычисляется путем сопоставления совместной встречаемости данного слова со словами *отличный (excellent)* и *плохой (poor)* в данной предметной области. Полученная оценочная направленность слов используется для классификации отзывов на положительные и отрицательные.

В работе [4] выделение оценочных слов и определение их семантической направленности основано на синтаксических шаблонах и союзах между словами. Основное внимание уделяется союзам И, ИЛИ и НО. Предполагается, что, если два прилагательных связаны союзами И или ИЛИ, то они оба являются или не являются оценочными, а так же одинаково семантически направлены. В случае союза НО, семантическое направление различается. Основываясь на этом принципе, был построен классификатор, определяющий семантическую направленность множеств прилагательных, работающий с точностью 92 %.

В работе [7] представлена система OPINE. Система служит для извлечения из отзывов разных атрибутов представленных продуктов, а также оценок по ним. OPINE выделяет следующие атрибуты продукта: свойства продукта, части продукта, атрибуты частей продукта, связанные сущности, свойства и части связанных сущностей. Предполагается, что оценочные фразы появляются в непосредственной близости от атрибутов объекта. Для извлечения оценочных слов используется 10 правил, основанных на синтаксической структуре предложения. Определение семантической ориентации слов базируется на ряде факторов, включая употребление с союзами, учет словообразования, информации о синонимах и антонимах из WordNet.

В работе [2] производился поиск мнений, выраженных придаточными предложениями. Как один из факторов извлечения таких мнений использовалась характеристика относительной частотности слов в документах с предполагаемым большим количеством мнений (колонка редактора и письма читателей) и в документах с меньшим количеством мнений (новости и бизнес-публикации).

Особенностью предлагаемого нами подхода является то, что значительное количество потенциальных характеристик оценочных слов вычисляется на основе сочетания разных текстовых корпусов в рамках заданной предметной области.

3. Получение корпусов и характеристик слов

3.1. Подготовка входных данных

Для подготовки данных с сайта *www.imhonet.ru* были собраны тридцать тысяч отзывов пользователей по различным фильмам. Кроме того, для каждого отзыва была извлечена численная оценка (от одного до десяти) фильма пользователем. Этот корпус является основным для работы, назовем его *корпус мнений*.

Пример отзыва (1): *Неплохой фильм, главное не выключить его в начале, где он напоминает просто ужасную пародию на Адреналин. Ну а в целом в фильме есть, как и положительные (адреналиновые, захватывающие и интересные сцены) так и отрицательные (неоднозначный финал, не везде удачная режиссура) качества.*

Для формирования нейтральной коллекции, где концентрация мнений значительно меньше, с того же сайта были собраны двадцать тысяч описаний фильмов. Назовем этот корпус *корпусом описаний*.

Собранные тексты были обработаны программой морфологического анализа и получен список лемм с информацией о части речи.

Для работы также использовался список лемм, с информацией об их встречаемости в новостном корпусе размером в один миллион документов. Условно этот список назовем *новостным корпусом*.

3.2. Составление корпуса с более высокой концентрацией оценочных слов

Было высказано предположение, что можно выделить некоторые части корпуса мнений, в которых концентрация оценочных слов больше, а именно:

- Предложения, заканчивающиеся на «!»;
- Предложения, заканчивающиеся на «…»;
- Короткие предложения не более чем из 7 слов;
- Короткие отзывы, состоящие из одного предложения;
- Предложения, содержащие слово «фильм» без других существительных.

Условно назовем это корпус — *малый корпус*. Его размер примерно в 2,5 раза меньше чем у *корпуса мнений*.

3.3. Предлагаемые характеристики

Для выделения качественного списка оценочных слов был предложен набор различных характеристик. Для подсчёта были выбраны следующие характеристики:

- Частотность.
- Количество документов, в которых встречается слово.
- Странность.
- TFIDF.
- Отклонение от средней оценки.
- Частотность слов, употребляемых с большой буквы (в корпусе мнений).

Остановимся более подробно на каждой из них.

3.3.1. Частотность

Частотность вычисляется как число появления слова во всем корпусе. Далее все слова упорядочиваются по убыванию частоты встречаемости.

3.3.2. Странность

Для подсчета характеристики странности необходимо два корпуса, один содержащий мнения, другой контрастный. Идея в том, что слова, которые несут оценки, будут «странными» в контексте контрастного корпуса [1], Сама характеристика вычисляется так:

$$\text{Странность} = (FRL/FRC)/(FRLC/FRCC)$$

FRL — частотность леммы в исследуемой коллекции.
FRC — число словоупотреблений во всей исследуемой коллекции.
FRLC — частотность леммы в контрастной коллекции.
FRCC — число словоупотреблений в контрастной коллекции.

Вместо частотности можно использовать количество документов, в котором встретилось слово.

3.3.3. TFIDF

Характеристика TFIDF хорошо известна в информационном поиске. Обычно она вычисляется на основе частотности некоторого слова в отдельном документе и в коллекции в целом. Мы подсчитываем эту характеристику на основе целых корпусов, тем самым также выявляем слова, которые «вдруг» повышают свою относительную частотность относительно другого корпуса.

Существует довольно большое количество способов подсчёта характеристики TFIDF, мы используем формулу из работы [9].

$$TFIDF(l) = \beta + (1 - \beta) * tf(l) * idf(l)$$

tf(l) — частота леммы l в корпусе с мнениями.
Idf(l) = log((|c|+0,5)/df(l))/log(|c|+1) — фактическая форма штрафования часто используемых в колекции слов.
df(l) — количество документов в контрастной коллекции, где встречалась лемма l.
β = 0,4.
|c| — количество документов в контрастной коллекции.

3.3.4. Отклонение от средней оценки

Как уже упоминалось, для каждого собранного текста мнения, сохранялась еще и числовая оценка (от одного до десяти), поставленная пользователем. Суть данной характеристики состоит в том, чтобы для каждого слова посчитать его среднюю оценку (т, е, взять оценки тех мнений, где оно встретилось, и разделить на количество словоупотреблений) и вычислить модуль разности со средней оценкой всего корпуса. Таким образом, мы получаем суммарную оценочную ориентацию этого слова.

$$dev(l) = \left| \frac{\sum_{i=1}^{n} m_i k_i}{k} - \frac{\sum_{i=1}^{n} m_i}{n} \right|$$

$$\sum_{i=1}^{n} k_i = k$$

l — рассматриваемая лемма.
n — общее количество отзывов.
m_i — оценка i-го отзыва.
k_i — число словоупотреблений леммы в i-ом отзыве (если не употребляется, тогда 0).

3.3.5. Частотность слов употребляемых с большой буквы

Суть этой характеристики в том, что имена собственные обычно не являются оценочными словами. Поэтому мы подсчитываем, сколько раз каждое слово употреблялось с большой буквы и при этом не находилось в начале текста или в начале предложения.

3.4. Комбинации характеристик и корпусов

Для экспериментов были взяты первые десять тысяч слов по частотности, и вся дальнейшая работа проводилась с ними. Слова были разделены на прилагательные и неприлагательные. Смысл такого разделения состоит в том, что многие исследователи указывали, что большинство оценочных слов являются прилагательными, и оценка качества нашего подхода на них, представляет отдельный интерес. В неприлагательные входят: существительные, глаголы и наречия. Все характеристики считались отдельно по этим двум категориям.

Таким образом, получаются такие комбинации характеристик и корпусов:

- TFIDF по парам корпусов: *малый-новости, малый-описания, мнения-новости, мнения-описания*;
- Странность по парам корпусов: *мнения-новости* и *мнения-описания* по количеству документов, *малый-описания* и *мнения-описания* по частотности;
- Отклонение от средней оценки;
- Частота по корпусу мнений и *малому корпусу*;
- Количество документов, в которых встречается слово в *корпусе мнений*;
- Частотность слов употребляемых с большой буквы в *корпусе мнений*;

Кроме этого, отдельно для *корпуса описаний* были посчитаны характеристики: частотность, количество документов, странность *описания-новости* по количеству документов и TFIDF по *корпусам описания-новости*.

Таким образом, для каждой леммы получается 17 признаков.

4. Оценка качества и комбинирование полученных списков

4.1. Метрика оценки качества

Для оценки качества получаемых списков слов мы использовали метрики оценки качества, применяемые для информационного поиска [10], В данной работе будут использоваться три метрики: точность, полнота и F-мера.

4.2. Точность (precision)

Точность вычисляется как отношение количества оценочных слов к общему количеству слов в списке:

$$P=a/(a+b)$$

Здесь a — количество слов, которые являются оценочными; b — количество слов, которые не являются оценочными. Например, если точность равна 50 %, то это значит, что в рассматриваемом списке половина слов — оценочные и половина — неоценочные.

4.3. Полнота (recall)

Полнота вычисляется как отношение найденных оценочных слов к общему количеству оценочных слов:

$$R=a/(a+c)$$

Здесь a — количество слов, которые являются оценочными; c — количество слов, которые являются оценочными, но не найдены. Например, если полнота равна 50 %, то это значит, что половина оценочных слов не найдена.

4.4. F-мера

F-мера часто используется как единая метрика, объединяющая метрики полноты и точности в одну метрику. F-мера вычисляется по формуле:

$$F=2PR/(P+R)$$

4.5. Разметка

Для оценки качества работы алгоритмов необходимо эталонное множество оценочных слов. Изначально была опробована идея построения эталонного множества по общезначимому списку оценочных слов. Были взяты около пятисот слов, но оценки качества, сделанные с помощью этого множества, не соответствовали действительности. Характеристики получались несоответствующие реальному положению дел, поскольку с помощью общезначимого списка нельзя получить слова, которые зависят от предметной области, жаргонные слова и некоторые другие. Поэтому было решено взять исходный десятитысячный список слов и вручную разметить в нем оценочные слова.

При разметке оценочных слов выяснилось, что во многих случаях нельзя сделать однозначный вывод о том, является ли слово оценочным, поскольку иногда слово является оценочным в другой области, но не является оценочным в рабочей области. Многие слова могли употребляться как в оценочном, так и неоценочном смысле при обсуждении фильмов. Поэтому было принято правило, что слово размечалось как оценочное, если можно представить какое-либо оценочное суждение по отношениям к фильмам и их атрибутам. Кроме того, разметка делалась обоими авторами работы.

В результате разметки получился список оценочных слов размером три тысячи двести слов (1262 прилагательных, 296 наречия, 857 существительных, 785 глаголов).

4.6. Результаты по отдельным характеристикам

Приведем результаты, полученные по каждой характеристике при подсчете среди первой тысячи слов:

Таблица 1. Прилагательные

Характеристика	Коллекция	Точность %
TFIDF	малый-новости	60,7
TFIDF	малый-описания	59,4
TFIDF	мнения-новости	60,0
TFIDF	мнения-описания	58,7
Странность	мнения-новости (количество документов)	64,0
Странность	мнения-описания (количество документов)	61,8
Странность	малый-описания (частотность)	60,7
Странность	мнения-описания (частотность)	61,4
Отклонение от оценки		56,3
Частотность	мнения	57,4
Частотность	малый	58,2
Количество документов	мнения	58,7

Таблица 2. Неприлагательные

Характеристика	Коллекция	Точность %
TFIDF	малый-новости	25,9
TFIDF	малый-описания	25,3
TFIDF	мнения-новости	23,2
TFIDF	мнения-описания	21,0
Странность	мнения-новости (Количество документов)	41,7
Странность	мнения-описания (Количество документов)	39,2
Странность	малый-описания (Частотность)	40,5
Странность	мнения-описания (Частотность)	38,2
Отклонение от оценки		30,6
Частотность	мнения	18,4
Частотность	малый	21,4
Количество документов	мнения	19,2

4.7. Машинное обучение

Имея для каждого слова набор характеристик, можно построить классификатор для автоматического разделения слов на оценочные и неоценочные. Для классификации использовалась свободно распространяемая система Rapid Miner [12], В данной работе использовались следующие алгоритмы:

- Метод k ближайших соседей (kNN)
- «Наивный» байесовский классификатор (Naïve Bayes)
- Персептрон (Perceptron)
- Нейронная сеть (2х и 3х-слойная)
- Логистическая регрессия (Logistic Regression)
- Метод опорных векторов (SVM стандартный и с радиальной ядровой функцией)

Оценки качества и подбор параметров алгоритмов производился с помощью кросс-валидации. Кроме того, воспользовавшись байесовским подходом к теории вероятностей, можно получить «вероятность» принадлежности объекта к классу оценочных слов. Если отсортировать слова по значению этой «вероятности», то можно узнать количество оценочных слов в первой тысяче слов списка.

4.8. Результаты классификации

Приведем результаты классифицирования для прилагательных и неприлагательных.

Таблица 3. Прилагательные

Алгоритм	Precision	Recall	F
kNN	63,98	70,75	67,17
Naïve Bayes	73,90	20,69	32,29
Perceptron	59,30	94,34	72,76
Neural Net(2 layers)	65,51	78,95	71,08
Neural Net(3 layers)	66,01	75,29	69,39
Logistic Regression	67,77	68,63	68,09
SVM	63,32	74,72	67,54

Таблица 4. Неприлагательные

Алгоритм	Precision	Recall	F
kNN	34,67	34,50	34,59
Naïve Bayes	28,44	88,39	42,56
Perceptron	53,48	5,88	10,39
Neural Net(2 layers)	38,22	28,32	32,52
Neural Net(3 layers)	55,90	14,90	23,19
Logistic Regression	58,70	9,18	15,84
SVM	37,15	27,00	31,27

В результате классификации: для прилагательных наилучшее значение F-меры получилось равным 72,76 % с использованием персептрона, для неприлагательных соответственно 42,56 % при классификации «наивным» байесовским алгоритмом.

Отдельный интерес представляет вычисление точности для первой тысячи слов, взятых от начала

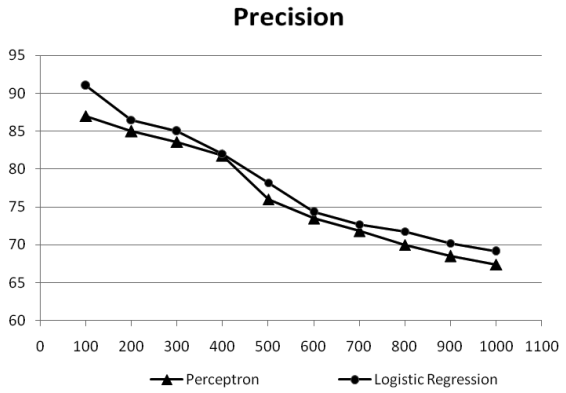


Рис. 1. Зависимость Точности от числа слов взятых сначала списка (прилагательные)

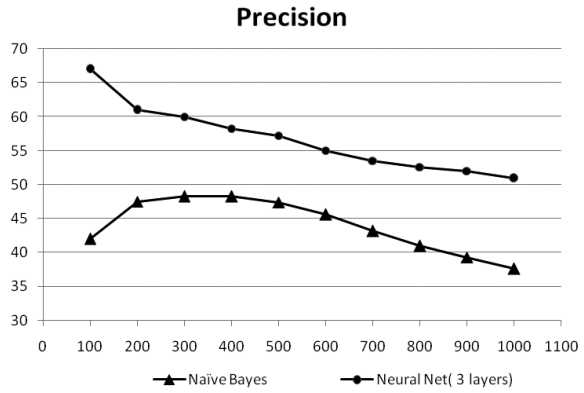


Рис. 2. Зависимость Точности от числа слов взятых сначала списка (неприлагательные)

списка, отсортированного по значению «вероятности» принадлежности объекта к классу оценочных слов. Эти показатели могут быть полезны для дальнейшего автоматического использования классифицированных слов. Наилучшее значение точности получились: для прилагательных при использовании логистической регрессии 69,1 %, для неприлагательных при использовании 3-х слойной нейронной сети 50,9 %, Интересным также показалось изобразить зависимость точности от количества слов взятых от начала. Для этого были взяты лучшие алгоритмы по точности на первой тысяче и лучшие по F-мере. Полученные результаты представлены в виде графиков.

Таким образом, удалось получить рост качества полученных списков на первой тысяче слов (по сравнению со списками по характеристикам) по точности для прилагательных — на 8,28 %, для неприлагательных — на 20,6 %.

В качестве примера приведем первые десять слов из лучших двух списков: по прилагательным и неприлагательным:

позитивный	*пересматривать*
отличный	*простой*
интересный	*тягомотина*
замечательный	*высосанный*
затянутый	*хавать*
смешной	*плоско*
добрый	*наигранно*
обалденный	*фигня*
предсказуемый	*блин*
потрясающий	*отвратительно*

5. Заключение

В работе мы показали, что путем извлечения нескольких корпусов заданной предметной области и вычисления с их помощью нескольких характеристик слов можно автоматически получить достаточно качественные списки оценочных слов. В качестве дальнейшей работы мы планируем добавить число характеристик слов, на основе которых можно улучшить качество выделения оценочных слов, а также предполагается оценить устойчивость предложенной технологии для другой предметной области.

Литература

1. *Ahmad K., Gillam L., Tostevin L.* University of Surrey participation in Trec8: Weirdness indexing for logical documents extrapolation and retrieval // In the Proceedings of Eigth Text Retrieval Conference (Trec-8), 1999.
2. *Bethard S., Yu H., Thornton A., Hatzivassiloglou V., Jurafsky D.* Automatic Extraction of Opinion Propositions and their Holders // AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications, 2004.
3. *Esuli A., Sebastiani F.* Determining the Semantic Orientation of Terms through Gloss Classification // Conference of Information and Knowledge Management, 2005.
4. *Hatzivassiloglou V., McKeown K.* Predicting the Semantic Orientation of Adjectives // ACL, 1997. P. 174–181
5. *Hu M., Liu B.* Mining and Summarizing Customer Reviews // KDD, 2004.
6. *Pang B., Lee L.* Opinion mining and sentiment analysis // Foundations and Trends® in Information Retrieval, Now Publishers, 2008
7. *Popescu A., Etzioni O.* Extracting Product Features and Opinions from Reviews // EMNLP, 2005.
8. *Turney P. D.* Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews // ACL, 2002, 417-424
9. *Агеев М. С., Добров Б. В., Лукашевич Н. В., Сидоров А. В.* Экспериментальные алгоритмы поиска/классификации и сравнение с «basic line» // РОМИП , 2004.
10. *Агеев М., Кураленок И., Некрестьянов И.* // Петрозаводск: Российский семинар по Оценке Методов Информационного Поиска (РОМИП 2009), 2009.
11. *Ермаков А. Е.* Извлечение знаний из текста и их обработка: состояние и перспективы // Информационные технологии, 2009.
12. *The RapidMiner toolset* // http://rapid-i.com

Исследование поискового спама, размещаемого посредством ссылочных брокеров

Research of web spam placed by link brokers

Шарапов Р. В. (info@vanta.ru), **Шарапова Е. В.** (mivlgu@mail.ru)

Муромский институт (филиал)
Владимирского государственного университета

В статье рассматриваются характеристики ссылочного спама, размещаемого ссылочными брокерами. Исследуется время жизни и ротация ссылок, анализируется тематическая близость ссылок и страниц.

Введение

Последнее десятилетие ознаменовалось бурным развитием глобальной сети Интернет и поисковых систем, позволяющих искать информацию в ней. Стремясь повысить качество поиска, поисковые системы стали использовать дополнительные сведения о документах, в том числе ссылки на них.

Ссылки используются для более эффективного ранжирования результатов поиска. В основе этого лежит постулат о том, что ссылка является воплощением желания поделиться полезной информацией с другими людьми, своего рода голосом за ресурс, на который ведет ссылка. Поэтому сайт, на который ведет много ссылок, вероятно, будет более полезен и интересен пользователям, чем сайт, на который никто не ссылается. Кроме того, ссылки с известных и популярных ресурсов считаются более весомыми, чем с никому не известных сайтов. Все это используется современными алгоритмами поисковых систем (PageRank, HITS, индекс цитирования), чтобы предоставлять пользователям наиболее нужную и полезную информацию по поисковым запросам.

В то же время, использование поисковыми системами ссылок привело к возникновению нового вида поискового спама, получившего название ссылочный спам [5]. Ссылочный спам заключается в формировании ссылочных структур, способных повлиять на алгоритмы работы поисковых систем с целью достижения более высоких позиций в результатах поиска по пользовательским запросам.

Ссылочный спам проявляется в размещении большого числа ссылок на сайтах с возможностью простого добавления информации (форумах, гостевых книгах, комментариях в блогах и т. д.). Такие ссылки предназначены в первую очередь для поисковых систем, а не для человека. В результате набираются искусственные «голоса» в пользу сайтов, на которые ведут эти спам-ссылки и сайты начинают лучше «искаться» поисковыми системами, оттесняя качественным и интересные ресурсы на второй план.

Еще большей проблемой являются системы пакетной покупки ссылок через ссылочных (рекламных) брокеров (таких как MainLink.ru, Xap.ru, Sape.ru, LinkFeed.ru, SetLinks.ru, Clx.ru и т. д.). Такие системы могут размещать ссылки на сотнях миллионов страниц. Массовое появление ссылок, размещаемых ссылочными брокерами, может оказать существенное влияние на алгоритмы поисковых систем [9]. Несмотря на то, что такие ссылки позиционируются как рекламные, считать их таковыми нельзя. Ссылки размещаются в неприметных местах страницы, чаще всего в самом ее низу, отображаются мелким шрифтом. Таким образом, функцию рекламы они выполнять не могут, так как пользователи такие ссылки просто не замечают. Следовательно, функцией ссылок, размещаемых ссылочными брокерами, является именно ссылочный спам.

1. Текущее состояние проблемы

Вопросам изучения ссылочного спама посвящено немало работ. Достаточно подробные обзоры состояния проблемы приведены нами в [10, 11].

Ряд работ посвящено изучению ферм ссылок и борьбе с ними. Например, в работе [8] предла-

гается анализировать вэб-граф для определения ссылочного спама. Проводится анализ входящих и исходящих ссылок сайтов, исследуется их пересечение. Рассматривается влияние ссылочного спама на алгоритм HITS.

В работе [3] проводится статистический анализ автоматически сгенерированных страниц со спамом. Авторы рассматривают отклонения от нормального распределения различных свойств страниц, включая имена сайтов, IP-адреса, входящие и исходящие ссылки, содержание страницы и норму изменения.

В [6] рассматриваются различные характеристики страницы (число слов на странице и в заголовке, длина слов, процент видимого текста и т. д.). Даются сведения о процентном содержании поискового спама в различных доменных зонах. Проводится сравнение выявленных характеристик с их распределением на «обычных» страницах, что способствует выявлению страниц, содержащих спам.

В работе [1] подробно анализируются ссылочные структуры, образующие вэб-граф. Исследуются различные характеристики, способствующие обнаружению ссылочного спама.

В работе [2] делается попытка определять ссылочный спам («непотистский» спам). Для решения задачи используется дерево решений C4.5. Авторы рассматривают 75 свойств, используемых для классификации. Эти свойства позволяют определять: совпадение заголовка и описания страницы, описание пересекается с текстом страницы, совпадение имен хостов, совпадение доменов, совпадение адресов страниц без доменов, совпадение некоторых частей IP адресов, одинаковые контактные E-mail домены и т. д.

В работе [7] рассматриваются две группы свойств, характеризующих ссылочный спам (для его обнаружения) — связанные с содержанием и со ссылочной структурой. К первой группе относятся: число слов на странице, средняя длина слов на странице, процент слов из списка популярных слов, процент видимого содержания страницы, число слов в заголовке страницы и т. д. Во второй группе относятся: процент страниц на наиболее популярном уровне, число входящих ссылок на страницу, число исходящих ссылок на страницу, отношение числа входящих и исходящих ссылок, число ссылок с главных страниц, процент входящих ссылок на наиболее популярные страницы, процент исходящих ссылок на наиболее популярные страницы, перекрестные ссылки на страницу, средний уровень страниц на сайте и т. д.

В [4] рассматривается понятие массы спама, меры воздействия спам-ссылок на ранг страницы. Рассматриваются вопросы оценки массы спама. Для определения спама активно используется ссылочная структура вэб-графа.

Несмотря на все разнообразие работ, подробного исследования ссылок, размещаемых с использованием ссылочных брокеров, не проводилось. Интерес представляет исследование таких ссылок с точки зрения их динамики и содержания, выявление свойств, способных помочь в борьбе с ними.

Цель нашего исследования — изучить характеристики ссылок, размещаемых с помощью ссылочных брокеров. В первую очередь нас интересовали динамические характеристики ссылок — как долго присутствуют ссылки на станицах, как часто они заменяются на новые ссылки и т. д. Кроме того, исследовалась тематическая близость ссылок, размещаемых ссылочными брокерами, и страниц, на которых они размещаются.

2. Источники данных

В качестве объекта исследования были выбраны 10 сайтов, размещающих ссылочный спам с использованием ссылочных брокеров. Сайты ежедневно сканировались в течение 7 месяцев (начиная с июня 2009 г.). Общее число сканируемых страниц составило около 5000 (число страниц менялось в связи с изменениями сайтов). На сайтах было размещено ежедневно около 5500 ссылок. Под «ссылкам» в данной работе мы будем понимать исключительно спам-ссылки, размещаемые с использованием ссылочных брокеров, исключая из рассмотрения естественные ссылки, также присутствующие на сайтах. Информация о факте размещения и месте расположения ссылок были предоставлены владельцами сайтов.

Сайты состояли из различного количества страниц — от 20 до более 2000, имели различную тематику (история, спорт, кино, мультфильмы, знаменитости/актеры, здоровье, музыка, мобильные телефоны, интернет-магазин и бизнес сайт). В период исследования основные показатели сайтов (число страниц, тематика, индекс цитируемости, PageRank и т. д.) не изменялись. По этой причине, влияние этих показателей на размещение ссылок в разные периоды времени, можно считать минимальным. Таким образом, процесс размещения ссылок на исследуемых сайтах, можно считать естественным.

Анализ полученных данных позволил выявить основные характеристики и особенности спам-ссылок, а также показатели, характеризующие ссылки, размещаемые посредством брокеров.

3. Ротация спам-ссылок

Коэффициент ротации ссылок (K_r) представляет собой отношение общего числа спам-ссылок за период исследования L_7 (7 месяцев) к числу ссылок, размещенных в настоящее время L_1:

Таблица 1. Статистика по размещению спам-ссылок

Сайт	Страниц P	Ссылок за 7 месяцев L_7	Ссылок сейчас L_1	Коэффициент ротации K_r	Коэффициент ротации в месяц K_{rm}
Сайт об истории	223	3162	1030	3,07	0,44
Сайт о мультфильмах	22	327	144	2,27	0,32
Сайт об актере	58	780	270	2,89	0,41
Сайт о спорте	110	3474	843	4,12	0,59
Сайт о здоровье	163	2252	1077	2,09	0,30
Бизнес сайт	86	1552	393	3,95	0,56
Сайт о музыке	169	1980	775	2,55	0,36
Сайт о телефонах	1322	1289	458	2,81	0,40
Сайт о кино	2316	3201	374	8,56	1,22
Интернет магазин	423	496	112	4,43	0,63
Всего	**4892**	**18513**	**5476**	**3,38**	**0,48**

$$K_r = L_7 / L_1$$

Коэффициент ротации спам-ссылок за месяц (K_{rm}) можно вычислить, разделив коэффициент K_r на количество месяцев, в течение которых проводились исследования:

$$K_{rm} = K_r / 7$$

Как можно заметить (таблица 1), коэффициент ротации спам-ссылок K_r изменяется в диапазоне от 2,09 до 8,56. Это означает, что за семь месяцев ссылки меняются от 2 до 8 раз. Среднее значение коэффициента ротации спам-ссылок составило 3,38. Аналогично, коэффициент ротации спам-ссылок в месяц K_{rm} меняется от 0,30 до 1,22, при среднем значении в 0,48.

4. Тематическая близость спам-ссылок и сайта

Анализ тематики ссылок, размещаемых с помощью ссылочных брокеров, дал также интересные результаты.

Тематическая ссылка — ссылка, тематика которой совпадает или близка к тематике страницы, на которой она размещается.

Для определения тематической близости была использована методика, применявшаяся нами в [10]. Среди всего числа ссылок L_1 (5476) количество тематических ссылок T оказалось достаточно небольшим — всего 242. В связи с тем, что распределение тематических ссылок по сайтам сильно отличается, интерес представляет относительный показатель — процент тематических ссылок T_{link}, вычисляемый по формуле:

$$T_{link} = T / L1 \times 100\ \%$$

Процент тематических ссылок изменяется в диапазоне от 0,7 до 10,6 %. Среднее значение T_{link} составило 4,4 %. Таким образом, в среднем только одна из 22 ссылок, размещаемых с использованием ссылочных брокеров, имеет тематику, совпадающую или близкую с тематикой сайтов.

Таблица 2. Количество тематических ссылок

Сайт	Ссылок L_1	Число тематических ссылок T	% тематических ссылок T_{link}
Сайт об истории	1030	7	0,7
Сайт о мультфильмах	144	7	4,8
Сайт об актере	270	14	5,2
Сайт о спорте	843	33	3,9
Сайт о здоровье	1077	82	7,6
Бизнес сайт	393	42	10,6
Сайт о музыке	775	4	0,5
Сайт о телефонах	458	20	4,3
Сайт о кино	374	24	6,4
Интернет магазин	112	9	8,0
Всего	**5476**	**242**	**4,4**

5. Тематическая близость в группе спам-ссылок

Спам-ссылки могут размещаться на странице как по одной, как и группами. Расположение ссылок отличается на различных сайтах. Некоторые сайты не содержат ни одной одиночной ссылки, а большинство групп состоит из 4–8 ссылок, другие — содержат в основном одиночные ссылки, и лишь иногда группы из двух-трех ссылок. Тем не менее, анализ групп показал интересный результат. Из 1023 групп ссылок, только в 178 группах оказалось по одной тематической ссылке (17,4 % от количества групп ссы-

Таблица 3. Группы ссылок

Сайт	Страниц *P*	Одиночных ссылок	Одиночных тематических ссылок	Групп ссылок	Групп с 1 тематической ссылкой	Групп с 2 и более тематическими ссылками
Сайт об истории	223	1	0	222	7	0
Сайт о мультфильмах	22	0	0	22	7	0
Сайт об актере	58	0	0	56	12	1
Сайт о спорте	110	0	0	110	29	2
Сайт о здоровье	163	0	0	163	69	6
Бизнес сайт	86	2	0	84	28	6
Сайт о музыке	169	0	0	169	4	0
Сайт о телефонах	1322	271	11	102	9	0
Сайт о кино	2316	120	12	73	10	1
Интернет магазин	423	49	6	22	3	0
Всего	**4892**	**443**	**29**	**1023**	**178**	**16**

лок), в 16 группах — по две и более тематических ссылок (1,6 %). Из 443 одиночных ссылок, только 29 оказались тематическими, что составляет всего 6,5 % от числа одиночных ссылок (таблица 3).

Таким образом, показатель тематической близость является отличительной чертой ссылочного спама, размещаемого ссылочными брокерами. Ссылки различаются по тематике как между собой (при размещении в группах), так с содержанием страницы, где они расположены. При этом, различие в тематике — колоссальное. Практически все ссылки имеют совершенно другую тематику. Приведем пример ссылок, размещаемых в период исследования на странице с биографией известного американского актера:

(23) аренда погрузчика от фирмы
(24) Оптимизация сайта seo поисковое продвижение сайтов сайт seo-studio.
(25) Триал спорт теннисный стол спорт инвентарь маты.
(26) новый коттедж готовые коттеджи
(27) **Скачать фильмы бесплатно**
(28) окна от производителя
(29) Метизы усовершенствованные. Метизы фильтры. Метизы классные. метизы.
(30) купить грунт
(31) Курсы менеджеров, курсы pr менеджеров.
(32) Костюм деда мороза и снегурочки. Заказывать Деда Мороза и Снегурочку.
(33) Wmz sms обменяй с гарантией. Wmr wme обмен дорого.
(34) tehsklad.ru предлагает пилы Makita
(35) Массовая рассылка смс от 1157. Рассылка смс от 1054.
(36) Банки переводов денег. Перевод денег с карты на карту лимит суммы альфа банк.
(37) Автомобили Тула, продажа авто Тула. Продажа б/у авто в городе Тула.
(38) Iso 9000, iso 9001 2008. Международного стандарта iso 9001 2008.
(39) интернет магазин часов копии.

Как можно заметить, только ссылка «Скачать фильмы бесплатно» (5) имеет хоть какое-то отношение к странице с биографией актера (и фильмы и актер связаны с кино). Все остальные ссылки не имеют ничего общего со страницей, и к тому же вряд ли будут интересны пользователям. Это является прямым доказательством того, что ссылки, размещаемые посредством ссылочных брокеров, являются именно ссылочным спамом и не предназначены для пользователей.

6. Время жизни спам-ссылок

Время жизни ссылки (D_{link}) — это период времени, в течение которого ссылка была размещена на странице (до момента ее удаления). Надо заметить, что некоторые ссылки могут кратковременно исчезать со страниц, а затем вновь появляться на них. В этом случае, ссылка считалась удаленной, если она не появлялась вновь в течение 10 суток с момента исчезновения.

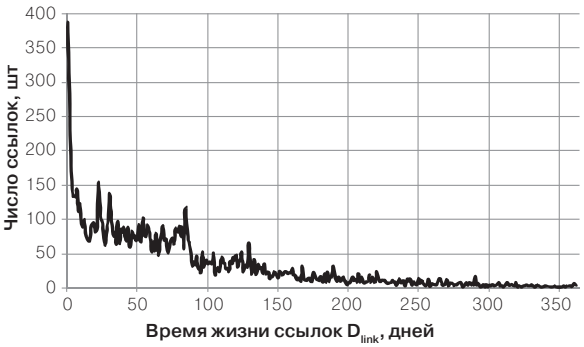


Рис. 1. График распределения времени жизни ссылок

На основании собранной статистики было получено распределение ссылок по времени жизни (количество ссылок, существовавших один, два, три и т. д. дней). На рисунке 1 показано распределение

времени жизни ссылок за 1 год. Число ссылок, имеющих время жизни больше одного года, продолжает уменьшаться, и к концу второго года сокращается до 1–2 штук.

Таблица 4. Распределение ссылок по времени жизни (месяцев)

Период	Процент ссылок, %
1 месяц	30,619
2 месяца	20,283
3 месяца	17,420
4 месяца	8,532
5 месяцев	7,349
6 месяцев	4,254
7 месяцев	3,252
8 месяцев	2,458
9 месяцев	1,746
10 месяцев	1,291
11 месяцев	0,786
12 месяцев	0,397
13 месяцев	0,546
14 месяцев	0,215
15 месяцев	0,223
16 месяцев	0,207
17 месяцев	0,232
18 месяцев	0,066
19 месяцев	0,074
20 месяцев	0,050

Рассмотрим процентный состав времени жизни ссылок, сгруппированных по месяцам. Как можно заметить, подавляющее число ссылок (более 50 %) существует не более 2 месяцев. Практически 90 % ссылок имеют время жизни не более 6 месяцев.

Таким образом, большинство ссылок имеют достаточно небольшое время жизни. Кроме того, ссылки, размещенные на одной странице (группой), также имеют разное время жизни. Поэтому, можно наблюдать ситуацию, когда, скажем, первая и третья ссылки в группе остаются неизменными, а вторая и четвертая ссылка успевают измениться несколько раз. Такие несбалансированные группы являются явным признаком ссылочного спама, размещаемого с использованием ссылочных брокеров.

7. Выводы

Таким образом, анализ ссылок, размещаемых с использованием ссылочных брокеров, показал, что они действительно предназначены для спама и не несут полезной информации для пользователей. Кроме того, практика показала достаточно невысокое время жизни таких ссылок. По этой причине ссылки со временем жизни более 6 месяцев, скорее всего, будут предназначены для пользователей, а не для поисковых систем.

Выявленные характеристики ссылок могут использоваться как исходный материал в алгоритмах обнаружения ссылочного спама и нейтрализации его действия на поисковые системы.

Мы планируем использовать полученные сведения в разрабатываемых нами алгоритмах обнаружения ссылочного спама [10, 11]. Выявленные характеристики ссылочного спама, размещаемого с использованием ссылочных брокеров, позволят расширить число параметров алгоритмов, что будет способствовать более точному определению спам-ссылок.

Литература

1. *Becchetti L., Castillo C., Donato D., Leonardi S., Baeza-Yates R.* Link Analysis for Web Spam Detection. ACM Trans. Web 2, 1, 1–42, 2008
2. *Davison B. D.* Recognizing nepotistic links on the web. In AAAI-2000 Workshop on Artificial Intelligence for Web Search, Austin, TX, July 30 2000, p. 23–28.
3. *Fetterly D., Manasse M., Najork M.* Spam, damn spam, and statistics — Using statistical analysis to locate spam web pages. In Proceedings of the 7th International Workshop on the Web and Databases (WebDB), Paris, France, 2004.
4. *Gyongyi Z., Berkhin P., Garcia-Molina H., Pedersen J.* Link Spam Detection Based on Mass Estimation. In: 32nd International Conference on Very Large Data Bases (VLDB 2006), September 12–15, 2006, Seoul, Korea.
5. *Gyöngyi Z., Garcia-Molina H.* Web spam taxonomy. In First International Workshop on Adversarial Information Retrieval on the Web, 2005, Chiba, Japan.
6. *Ntoulas A., Najork M., Manasse M., Fetterly D.* Detecting spam web pages through content analysis. In Proceedings of the World Wide Web conference, Edinburgh, Scotland, May 2006, p. 83–92.
7. *Qingqing Gan, Torsten Suel.* Improving web spam classifiers using link structure. Proceedings in Third International Workshop on Adversarial Information Retrieval on the Web (AIRWeb '07), May 2007, Banff, Alberta, Canada
8. *Wu B., Davison B. D.* Identifying link farm pages. In Proceedings of the 14th International World Wide Web Conference (WWW), 2005.
9. *Шарапов Р. В., Шарапова Е. В.* Обнаружение ссылочного спама // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды Десятой Всероссийской научной конференции «RCDL'2008» (Дубна, Россия, 7–11 октября 2008 г.). Дубна: ОИЯИ, 2008. С. 191–196.
10. *Шарапов Р. В., Шарапова Е. В.* Алгоритм обнаружения ссылочного спама // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог 2009» (Бекасово, 27–31 мая 2009 г). М: РГГУ, 2009. Вып. 8 (15). С. 537–542.
11. *Шарапов Р. В., Шарапова Е. В.* Применение метода опорных векторов для обнаружения ссылочного спама // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды XI Всероссийской научной конференции «RCDL'2009» (Петрозаводск, Россия, 17–21 сентября 2009 г.) Петрозаводск: КарНЦ, 2009. С. 318–324.

Конструкция «праздник не в праздник» на фоне других конструкций

The peculiarities of the russian idiomatic construction «prazdnik ne v prazdnik»

Шеманаева О. Ю. (shemanaeva@yandex.ru)

Институт проблем передачи информации им. А. А. Харкевича РАН

Показано своеобразие русской конструкции типа *праздник не в праздник* на фоне других конструкций, похожих на нее по структуре и семантически. Эта статья является дополнением к инвентарю русских конструкций или синтаксических фразем — выражений некомпозициональной природы, находящихся между грамматикой и словарем.

1. Постановка задачи

Работа посвящена особому типу выражений русского языка вида **X не в X** — *праздник не в праздник*, которые находятся на языковой периферии — в сфере идиоматики, употребляются сравнительно редко, относятся к разговорной и в том числе к устаревшей речи. Они входят в более обширный класс конструкций с лексическими повторами, о которых см. такие работы, как [Шведова 2003], [Wierzbicka 1987], [Плунгян, Рахилина 1996], [Санников 2008], [Апресян В. 2010] — о родственных конструкциях *день в день, час в час, минута в минуту*, см. также [Мальцева 2006] и др.

Для понимания данного типа выражений особенно важно не сложение смыслов участников конструкции, а общий смысл, благодаря которому это выражение является не полностью связанным, но носители языка могут порождать новые высказывания, а смысл конструкции позволяет понять это выражение и с другими участниками. Мало зафиксировать самые частотные выражения такого рода во фразеологическом словаре (*праздник не в праздник, пир не в пир*), надо описать смысл целой конструкции вида **X не в X**.

О такого рода выражениях говорят в терминах грамматики конструкций [Fillmore et al. 1988], [Goldberg 1995], [Jackendoff 1997], см. также работы [Иорданская, Мельчук 2007], [Kopotev 2008], где они называются синтаксическими фраземами, а в Московской семантической школе эти конструкции получили название выражений «малого синтаксиса» [Иомдин 2003, 2006, 2008].

Мы рассмотрим принципы построения конструкции *праздник не в праздник* (выбор участников), синтаксическое окружение **X не в X** и общий смысл этой конструкции, благодаря которому введение в нее новых элементов воспринимается правильным образом и возможна языковая игра и наполнение известной схемы новыми лексическими единицами.

Изложение будет построено следующим образом: 1) более общая конструкция **(Z-y) Y в X** (*конференции ей в диво / в радость / в тягость / в новинку* и т. д.), в которой Y и X различны, 2) собственно конструкция с повтором (**Z-y) X не в X**, которая является подклассом более общей конструкции из пункта 1); структурные и семантические особенности конструкции **X не в X**; 3) Сравнение с похожими конструкциями того же семейства: **X в X** (имеется повтор, наличие предлога, отсутствие отрицания) и **X не X** (имеется повтор, наличие отрицания, отсутствие предлога).

В работе были использованы данные толковых и фразеологических словарей, а также Национального корпуса русского языка (www.ruscorpora.ru) и примеры, полученные в интернете с помощью точного поиска.

2. Конструкция Y в X как основание для конструкции с повтором X в X

Выражения вида Y в X *(не) в диковинку, (не) в диво, в новинку, в тягость, в радость, в охоту /*

в охотку, в укор, в досаду, (не) в обузу, (не) в горесть, в кайф (сленговый синоним *в радость / в удовольствие*) и *в труд* (синоним *в тягость / в обузу*) представляют собой сочетания предлога *В* с винительным падежом. В Синтаксическом словаре таким выражениям приписано следующее значение: «эмоционально-оценочная реакция, каузируемая наличием предмета, событием, состоянием или действием» [Золотова 2001: 167]:

(1) *…а сёстрам то в досаду было, и задумали они дело хитрое…* (С. Т. Аксаков)

Конструкции Y в X бывают как утвердительные, так и с отрицанием: *(не) в диковинку, (не) в новинку, (не) в диво, (не) в охоту / (не) в охотку*, *(не) в укор, (не) в упрек, (не) в досаду, (не) в обузу, (не) в горесть, (не) в кайф, (не) в труд* и др.

Два участника этой конструкции — *радость* и *труд* — чаще других образуют конструкции с повтором и отрицанием: *радость не в радость*, *труд не в труд*. Именно эти два смысла присутствуют в семантике конструкции X не в X.

3. Семантика конструкции X не в X

У этой конструкции обнаружено два значения: условно «не в радость» и «не в тягость», в зависимости от оценки говорящим события X. Одна и та же конструкция, таким образом, несет в себе определенную оценку события, но в зависимости от аргумента и оценки его говорящим, оценка будет положительной или отрицательной, ср. слово *дерзкий* в выражении *дерзкий мальчишка* ('непочтительный', отрицательная оценка) и *дерзкая атака* ('вызывающе смелая', возможна положительная оценка).

1) Значение «не в радость»

Говоря «**Z-у и X не в X**» (*мне и праздник не в праздник*), говорящий имеет в виду, что событие или процесс X не соответствует эталонному событию или процессу X, и тем самым не приносит человеку Z, эмоционально оценивающему эту ситуацию, никакой радости, никаких положительных эмоций. Другая интерпретация — X соответствует эталону, но по какой-то причине Z не реагирует на X так, как реагирует обычно или как принято реагировать.

Неудивительно, что в значении «не в радость», эталонный X должен обладать чем-то притягательным для говорящего, он должен быть заинтересован в X, у X-а часто положительная оценка (*жизнь, радость, праздник, рай, отпуск, отдых, бал, пир* и т. д.)

(2) *Благовещенье… и каждый должен обрадовать кого-то, а то* ***праздник не в праздник*** *будет.* [И. С. Шмелев. Лето Господне (1927–1944)]

(3) *Поймите, из-за ваших звонков* ***жизнь*** *моей жене* ***не в жизнь*** *и* ***радость не в радость****.* [Владимир Маканин. Ключарев и Алимушкин (1974)]

(4) *О люди! все похожи вы На прародительницу Еву: …Запретный плод вам подавай. А без того вам* ***рай не в рай*** (А.С. Пушкин. Евгений Онегин)

(5) *А какой это был забавник! Без него* ***свадьба не в свадьбу*** *была, скорей походила на похороны.* (Ш. Алейхем, пер. Б. Ивантер, Р. Рубиной)

(6) *Пойдем-ка к гостям, у нас без тебя и* ***пир не в пир****.* (А. Островский)

(7) *А для неверующего и* ***чудо не в чудо****; при нем хоть воскреси из мертвых человека, останови солнце, потряси землю, он подивится и останется неверующим* (Старец Анатолий Оптинский)

(8) *Нам* ***рассвет не в рассвет****, нас почти уже нет…* (А. Макаревич)

Плохие события, от которых не приходится ждать радости и удовольствия, не встречаются в этой конструкции: *???смерть не в смерть, ???увольнение не в увольнение*[1]*, ???развод не в развод*, и под.

2) Значение «не в тягость / не в труд»

Несмотря на плохие коннотации X, в ситуации в целом есть что-то, что помогает пренебрегать трудностями, связанными с X. Интересно, что конструкция в целом обладает энантиосемичным свойством иметь то отрицательную, как в пункте 1), то положительную оценку.

(9) *Всё ей было как будто забавно, весело,* ***лагерь не в лагерь****, хоть в комсомол вступай.* (А.Солженицын)

(10) *Очень хорошо подбегать к дереву и греться в такую погоду, прямо* ***дождь не в дождь*** *и* ***холод не в холод****.* (Интернет)

(11) *…с сожалением покидали дом, где царили весь, такой редкий в их жизни день, где* ***труд был не в труд****, в удовольствие и праздник.* (В. Астафьев)

(12) *Немощь пройдет; возвратится мужество, и опять* ***труд будет не в труд****. Так уже устроил нас Господь.* (Ф. Затворник)

[1] Ср., однако, *увольнение из армии,* которое может быть воспринято положительно (пример Л. Л. Иомдина).

Так, *труд* попадает во вторую группу, а *работа* — в первую, так как связано с времяпровождением, полезным и приятным для человека: *работа не в работу* означает, что человеку «не работается, не хочется работать, работа не клеится», а не то, что ему работать не трудно. Выражение *работа не в работу* имеет отрицательную оценку:

(13) *По-моему, шутить и смеяться надо, и русскому без этого* ***работа не в работу***. (Г.П. Блок)

(14) *…она писала мне, что поехала она туда напрасно, что ей и лыжи не в лыжи*[2]*, и* ***работа не в работу****, и тишина и лес ни к чему, что всюду перед ней этот деревянный загон, этот беспомощный и сильный человек, эта стража…* (Л. К. Чуковская)

4. Структура конструкции X не в X

4.1. Общий вид

Распространены следующие структурные типы данной конструкции (с факультативными элементами):

1) (Z-у) (и) X не в X (*без тебя мне праздник не в праздник*)

Усилительная частица *и* встречается факультативно, экспериенцер Z может быть выражен или не выражен. Прагматически часто встречается предложная группа **без Y-а: без Y-а (Z-у) (и) X не в X** Предложная группа «без Y-а» чаще расположена в препозиции к конструкции **X не в X,** ср. пример с постпозицией ?*мне праздник не в праздник без тебя*. **2) X не в X, (если) / (когда) / (коли) <устар.> / (как) <устар.>… (***праздник не в праздник, если тебя нет рядом***)**

(15) *Ведь, в сущности,* ***награда не в награду****, коли до нас не дошла и дано не то совсем, что назначено.*

Возможно дистантное расположение повторяющихся элементов конструкции «X» и «не в X» за счет распространения ИГ X или за счет вставки элементов (*без* Y-а) (Z-у) или за счет появления глагола-связки в будущем или прошедшем времени (*будет / был*), реже другой лексемы (*вышел, станет*).

(16) *Все трое бледны и угрюмы, / И* ***пир*** *<u>веселый им</u>* ***не в пир***. (А.С. Пушкин, Руслан и Людмила)

(17) *К нему приставили двенадцать слуг, и каждый держал его за привязанную к лапке шелковую ленточку. И* ***прогулка*** *<u>была ему</u>* ***не в прогулку***. (Г-Х. Андерсен)

Вернемся к порядку элементов внутри конструкции: препозиция группы «не в X» возможна в поэтических примерах, а вне поэзии нестандартна:

(18) *Маялся рабочий, а теперь себя буржуй помай. Спекулянтам всем* ***не в праздник — праздник*** *Май.* (В. Маяковский)

(19) *Но была* ***не в радость — радость****.* (Ф. Глинка)

4.2. Заинтересованный участник Z (субъект состояния)

Z может быть выражен местоимением: *мне / тебе / ему / ей / нам/ вам/ им*, может быть выражен полной именной группой:

(20) *Студенты ставили Гамлета, / и в этот день был* ***рай не в рай*** */ великой тени барда.* (В. Набоков)

4.3. Вариативность X

Эта область участников достаточно свободна, однако и здесь можно попытаться выделить некоторые если не запреты, то ограничения на использование. Наиболее естественные X — процессы, события, отрезки времени, мероприятия (а не артефакты, ср. маловероятность примера ??*шуба не в шубу,;* в то же время если артефакты рассматриваются не как физические объекты, а как события, связанные с ними, то такое употребление вполне допустимо (за названием артефакта стоит какой-то предикат): *лыжи не в лыжи* (имеется в виду *катание на лыжах)*, *книга не в книгу (чтение книги* или *написание книги)*) — можно разделить на следующие подклассы:

X — Названия праздников и похожих на праздники значимых периодов (*пост, масленица, Пасха, отпуск, отдых, …*)

(21) *Посоветуйте, пожалуйста, мне как успокоить кота на моё отсутствие, а то он жизни не даст в доме. А то и* ***отпуск не в отпуск****.* (Интернет)

(22) *Осе* ***отдых не в отдых*** *из-за Шуры — не понимаю, что с ним случилось.* (Н. Мандельштам)

X — названия деятельности человека, важной для него — игры: *спорт, футбол*, преферанс, …; **названия важных событий-мероприятий**: *бал, поход, танцы, свидание, концерт, ремонт, …*

[2] Отметим, что имеются в виду *лыжи* не как предмет, а лыжи как процесс: ‘катание на лыжах’.

(23) *Да еще вина тех поклонников бокса, которым непременно подавай нокауты — без них, мол,* ***спорт не в спорт****, «не зрелищно».* (Интернет)

(24) ***Урок не в урок*** *был, когда приходил отец Иван. Только в классе появится, закричим „ура“, безобразничать начинаем».* (П. В. Вавилов)

(25) *Даже женатый, он не пропускал ни одного хоровода, ни одной посиделки. Без него и* ***веселье*** *было* ***не в веселье****, и* ***пляс не в пляс****, и* ***игры не в игры****.* (Интернет)

(26) *И в первый праздничек соседи мчат к соседу, / Но* ***пир гостям — не в пир, беседа — не в беседу****, / И* ***преферанс — не в преферанс****! Им гипнотический подай скорей сеанс.* (Д. Бедный)

(27) *Иначе все пойдет вверх дном / До часа расставанья: / И сад не в сад, и дом не в дом, /* ***Свиданье — не в свиданье!*** (А. Ахматова)

X — Названия приемов пищи: *завтрак, обед, ужин* (в параллель к пиру)

Внутри этого класса возникает более дробная классификация в связи с типом Y:

1) Y — человек (хозяин, хозяйка)

(28) ***Обед не в обед****, когда хозяюшки нет / как хозяина нет.* (В. Даль)

(29) *…да почему его нет? И* ***обед не в обед****. Тогда уж к нему даже кого-нибудь и отправят депутатом проведать, что с ним, не заболел ли, не уехал ли?* [И. А. Гончаров. Обыкновенная история (1847)]

2) Y — еда или питье

(30) *Без каши и* ***обед не в обед;*** *и* ***обед не в обед****, коли хлеба нет.*

(31) *Считалось особым шиком, когда обеды готовил повар-француз Оливье, еще тогда прославившийся изобретенным им «салатом Оливье», без которого* ***обед не в обед*** *и тайну которого не открывал.* (В. Гиляровский, Москва и москвичи)

(32) *Без кислого блюда для рабочего* ***обед не в обед****.* [А. Н. Энгельгардт. Письма из деревни (1872–1887 гг.). Письмо первое (1878)]

(33) *Без спиртного и* ***ужин не в ужин****.* (В. Шаламов)

(34) *Не хочу овсянку по утрам, а вот без вредного кофе* ***завтрак не в завтрак****.* (Интернет)

3) Y — дополнительные условия, при которых можно наслаждаться таким мероприятием, как прием пищи

(35) *Благодарю вас. Но можно мне цветок в петлицу? Без цветка в петлице мне и* ***обед не в обед****.* (О. Уайльд. Как важно быть серьезным, пер. И. Кашкина)

(36) *Да, это дурная привычка, вдобавок страшно вредная, но если человек курит две пачки в день и ему это в кайф, то ему и* ***кофе не в кофе*** *без сигареты, а 5 часов в поезде — просто мученье.* (Интернет)

X — Названия времен суток: ***день, ночь***

(37) ***Ночь*** *ему* ***не в ночь****, ни сна, ни житья с ним, проклятым.* [Б. Л. Пастернак. Доктор Живаго (1945–1955)]

(38) *Мне* ***ночь не в ночь****, мне в ночь невмочь, когда тебя нету со мной.* (Н. Гянджеви, пер. К. Липскерова)

(39) *В голодной и больной неволе / И* ***день не в день****, и год не в год. / Когда же всколосится поле, / Вздохнет униженный народ?* (А. Блок)

X — Названия времен года: ***зима, осень***

(40) *Не было рябины — говорили, что и* ***осень не в осень****.* (Интернет)

(41) *Главное, мы уловили те несколько (буквально!) часов, когда можно было ее слепить! Отвели все-таки душу, а то как-то и* ***зима не в зиму*** *была.* (Интернет)

Конструкциям с повторами в русском языке, кроме монографии [Шведова 2003], уделено внимание в книге [Санников 2008]. Близки по структуре к рассматриваемой нами конструкции *праздник не в праздник* такие выражения с тождественными элементами, где второму элементу предшествует отрицательная частица *не*: *Я верил и не верил, пьян не пьян, а както дико смотрит, спал не спал, а вставай*, а так же тавтологические высказывания *бывают аварии и аварии*. В. З. Санников отмечает, что в тавтологических высказываниях такого рода требуется множественное, а не единственное число повторяющегося существительного, видимо, потому, что по смыслу эта конструкция является некоторым обобщением, общим заключением. Наоборот, в нашей конструкции *праздник не в праздник* предпочтительнее единственное число, выражающее отношение к конкретному событию, а не к праздникам вообще.

5. Родственные конструкции

5.1. Конструкция X в X

Конструкция *праздник ему был не в праздник* со значением несоответствия желаемому и несоответствия идеальному, нормальному представлению об X, практически не употребляется без отрицания. Однако у конструкции X в X, где второй, повторяющийся, компонент, также стоит в винительном падеже, есть другой смысл: пространственного или метафорического соответствия, при том, что X-ы принадлежат разным участникам одной и той же ситуации, ср.: *Их квартиры расположены дверь в дверь* или *Они живут душа в душу (*кроме *душа в душу*, в метафорическом (и в пространственном) значении встречается *рука в руку* и — сниженно — *рыло в рыло.)*, а также *они идут нога в ногу*. Для этой конструкции как раз, в отличие от *праздник не в праздник*, крайне неестественно отрицание предложной группы **в X**: **жить душа не в душу, *дома стоят окна не в окна*. Предикаты в данной конструкции — *жить, находиться (рядом / напротив), работать*. X в пространственном значении могут быть: *дверь, окно, окна, дом, забор, ворота, калитка, балкон, комната, стол, борт (*Заметим, что данная конструкция довольно близка по смыслу к пространственной конструкции ориентирования *лицом к лицу*, описанной в [Подлесская, Рахилина 2000] и к пространственной конструкции вида **X (им. п.) к X (дат. п.)** — *нос к носу)*.

(42) *Да, по-доброму соседствовали, прямо* ***стол в стол****, такие разные, казалось бы, просто не совместимые друг с другом личности* (Интернет)

Кроме этого, на периферии данной конструкции есть такие терминологические выражения, как *шов в шов, конец в конец* или *край в край (зашить рану, напечатать текст*), в которых имеется в виду точное соответствие пространственного положения одного шва / конца / края другому.

5.2. Конструкция X не X

Очень близкой по смыслу и структуре является конструкция с отрицанием, но без предлога: *праздник не праздник***.** На близость этих конструкций указывает и то, что в ряде примеров в сочинительной конструкции используются одновременно оба типа:

(43) *Без красного у русских и* ***бой не бой, праздник не в праздник****: уж коли кушака нет алого, зато нос клюквой.* [А. А. Бестужев-Марлинский. Вадимов (1834)]

Одним из значений этой конструкции является несоответствие норме, эталонному X, так же, как и в конструкции X не в X:

(44) *Без твоей песни нам и праздник — не праздник.* [Т. Г. Габбе. Город мастеров, или Сказка о двух горбунах (1943)]

(45) *Без театра, без подмостков мне жизнь — не жизнь.* [Владимир Брагин. В стране дремучих трав (1962)]

Характерным отличием нашей конструкции от этой конструкции без предлога *В*, с одним отрицанием, является отсутствие контраста, ср.: *праздник не в праздник* vs. *праздник не праздник, а так, просто гости*.

Впрочем, кроме значения, близкого значению конструкции **X не в X**, у конструкции с тавтологическим отрицанием без предлога В есть и свои особенные значения, например:

- условно-уступительная конструкция

(46) *— Да уж радость не радость, а пришли — угощай!* [В. И. Немирович-Данченко. Святые горы (1880)]

(47) *Весна не весна, а дорога испортилась и ее больше не будет…* [М. М. Пришвин. Дневники (1922)]

- неопределенность статуса X: скорее не X, но нечто похожее на X

(48) *Виделось мне, будто у нас на «Надежде» смотр не смотр, праздник не праздник, только народу кишмя кишит: генералов, адмиралов, штабства — видимо-невидимо.* [А. А. Бестужев-Марлинский. Фрегат «Надежда» (1833)]

(49) *Ну, знаешь, свадьба не свадьба, а какая-то чертовщина там происходит!* [Валентина Осеева. Динка прощается с детством (1969)]

6. Заключение

В заключение надо сказать, что в рамках данной статьи мы рассмотрели не все родственные конструкции, близкие по структуре и по смыслу к выражению **X не в X**. Это — задача дальнейшего исследования конструкций с повторами в русской разговорной и идиоматической речи.

Литература

1. *Fillmore Ch., Kay P. & O'Connor M.* Regularity and idiomaticity in grammatical construcions: the case of *let alone.* Language 63(3): 501–538. 1988.
2. *Goldberg A.* Constructions: A constructionist grammar approach to argument structure. Chicago: Chicago University Press, 1995.
3. *Jackendoff R.* Twisting the night away. Language, Vol.73, №3, 1997. 535–559.
4. *Kopotev M. Принципы синтаксической идиоматизации. Хельсинки, 2008.*
5. *Wierzbicka A.* Boys Will Be Boys: Radical Semantics vs. Radical Pragmatics // Language, 1987a, v.63, no.1, p.95–114. // 2003 (1991). *Cross-Cultural Pragmatics: The semantics of human interaction*. Berlin: Mouton de Gruyter. (Expanded second edition).
6. *Апресян В. Ю.* День в день, час в час, минута в минуту // Слово и язык, в печати.
7. *Золотова Г. А.* Синтаксический словарь. Репертуар элементарных единиц русского синтаксиса. М., УРСС, 2001.
8. *Иомдин Л. Л.* Большие проблемы малого синтаксиса. // Труды международной конференции по компьютерной лингвистике и интеллектуальным технологиям Диалог 2003. М.: Наука, 2003. — С. 216–222.
9. *Иомдин Л. Л.* Многозначные синтаксические фраземы: между лексикой и синтаксисом. // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог'2006» (Бекасово, 1–6 июня, 2006 г.). / Под ред. И. М. Кобозевой, А. С. Нариньяни, В. П. Селегея. — М.: Наука, 2006. — С. 202–206.
10. *Иомдин Л. Л.* В глубинах микросинтаксиса: один лексический класс синтаксических фразем. // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 4–8 июня, 2008 г.). Вып. 7(14). — М.: РГГУ, 2008. — С.178–184.
11. *Иорданская Л. Н., Мельчук И. А.* Смысл и сочетаемость в словаре // М.: Языки славянских культур, 2007.
12. *Мальцева В. С.* Конструкции с лексическими повторами в составе контрастивных предложений (на материале русского языка) // Третий международный симпозиум по языкам Европы и Северной и Центральной Азии (LENCA), 2006.
13. *Плунгян В. А., Рахилина Е. В.* «Тушат-тушат — не потушат»: грамматика одной глагольной конструкции // Змарзер В., Петрухина Е. П. (ред.) Исследования по глаголу в славянских языках: глагольная лексика с точки зрения симантики, словообразования, грамматики. М.: Филология, 1996.
14. *Подлесская В. И., Рахилина Е. В.* Лицом к лицу // ЛАЯ: Язык пространства. М., ЯРК, 2000.
15. *Санников В. З.* Русский синтаксис. М., 2008.
16. *Шведова Н. Ю.* Очерки по синтаксису русской разговорной речи. М.: Азбуковник, 2003 (1-е изд. — 1960).

Электронный двуязычный словарь метафор психологической сферы человека

Electronic bilingual dictionary of metaphors of human psychology sphere

Шиманская О. Ю. (shimans@tut.by)

Белорусский государственный педагогический университет им. Максима Танка, Минск, Беларусь

В статье описывается структура и методы создания электронного двуязычного словаря метафор психологической сферы человека (на базе субстантивов белорусского и английского языков).

1. Введение

Сопоставление метафор, репрезентирующих психологическую сферу человека, в белорусском и английском языках показывает, что метафорические обозначения различных феноменов психологии человека (психических процессов и состояний, свойств личности, психомоторики, поведения и общения) имеют много общего в обоих языках, а также характеризуются рядом специфических черт, обусловленных лингвистическими и этнокультурными факторами.

Практическим результатом сопоставления метафор (а также одним из средств обработки данных) стало создание электронного двуязычного словаря, который не только отражает семантику метафор, но и осуществляет поиск синонимов и эквивалентов, а также дает возможность сортировать метафоры относительно источника и цели мотивации.

Созданный на основе современных словарей литературного языка и выборки из художественных и публицистических текстов, словарь полезен как для решения актуальных проблем современной лексикологии, типологии метафорических моделей, теоретической и практической лексикографии, так и для совершенствования уровня языковой компетенции пользователей ресурса.

2. Общая характеристика словаря

Двуязычный словарь метафор психологической сферы человека является электронным ресурсом, отражающим результаты анализа и систематизации метафор внутреннего мира человека в белорусском и английском языках. Объект словарной систематизации — субстантивные метафоры, используемые для номинации и характеризации психических процессов, состояний и свойств личности.

Словарь ориентирован на широкий круг пользователей — филологов различного профиля, преподавателей, писателей, переводчиков, журналистов, психологов, всех, кто изучает белорусский и английский языки, интересуется образными средствами репрезентации внутреннего мира человека в языке. В словаре отражается:

- семантика метафор, репрезентирующих психологическую сферу человека;
- фонетические и морфемные варианты метафор;
- синонимические средства репрезентации в языке одного и того же денотата (с различной степенью близости метафорических значений);
- перевод каждой метафоры на английский (белорусский) язык с учетом контекста;
- потенциальные языковые метафоры, которые соответствуют регулярным метафорическим моделям и могут использоваться как при переводе, так и в качестве синонимов языковых метафор;
- средства метафорической репрезентации внутреннего мира человека (на основе тематического реестра определяемых понятий);
- характер денотации, непосредственный выбор предметов и явлений реальности для описания явлений внутреннего мира человека в белорусском и английском языках.

Следует отметить, что словарь ориентирован на пользователей, владеющих белорусским и (или) английским языком, и предполагается, что основное (неметафорическое) значение каждого словарного слова является знакомым пользователю, и потому объектом систематизации выступают именно мета-

форические значения. Для систематизации метафор относительно источника мотивации используются тематические группы, что позволяет установить область референции метафор, не перегружая словарь дополнительными толкованиями.

Основной реестр словаря составляют 1285 метафор — 632 белорусские метафоры (490 узуальных и 142 потенциальные языковые) и 653 английские метафоры (548 и 105 соответственно).

Источниками для словаря послужили материалы, выделенные путем сплошной выборки из различных словарей [1; 2; 3; 8; 9; 10; 11], а также метафоры из художественных и публицистических текстов, в том числе из Интернет-ресурсов.

Для систематизации данных была создана база данных в СУБД Microsoft Office Access 2003, что соответствует задачам словаря и делает его доступным практически каждому пользователю ПК, избавляя от необходимости установки дополнительных программ.

В настоящее время для словаря разработан только белорусский интерфейс, однако вскоре планируется создание английского интерфейса, а также добавление метафор из русского языка, что значительно расширит сферу применения данного ресурса.

3. Состав словаря и структура словарной статьи

Электронный ресурс содержит руководство пользователя, краткий теоретический экскурс, тестовые задания и БД «Словарь» и «Поиск».

БД «Словарь» позволяет пользователю выбрать интересующую его метафору из реестра, где английские и белорусские слова расположены в алфавитном порядке. При выборе слова на экран выводится электронная карточка, соотносимая со словарной статьей. В словаре можно быстро искать нужную карточку путем ввода слова в строку поиска или использования полосы прокрутки. Словарь может использоваться как толковый и синонимический словарь метафор в рамках одного языка, а также как переводной словарь для перевода метафор с белорусского на английский язык и наоборот.

Структура электронной словарной карточки следующая: заголовочное слово, толкование, примеры, синонимы (отмечены ○) и эквиваленты (отмечены ●). Раздел синонимов и эквивалентов имеет отсылочный статус, причем в будущем планируется его представление в виде гиперссылок на соответствующие словарные карточки.

Если метафора имеет два и более значений, то каждое толкование приводится с новой строки в последовательной нумерации. Иллюстрации, синонимические и эквивалентные варианты приводятся после толкования, к которому они относятся. Следует отметить, что наличие примера-иллюстрации для каждого метафорического значения является одной из доминант словаря, поскольку именно контекст ориентирует в основных правилах употребления метафор в речи. Для подбора наиболее близкого по значению синонима или эквивалента пользователю рекомендуется сопоставить контексты употребления выбранных метафор.

Потенциальные языковые метафоры, в том числе самостоятельные метафоры и отдельные метафорические значения, имеют помету *. В словарной статье могут помещаться грамматические и стилистические пометы, указания по сочетаемости метафор в контексте, что имеет особое значение при переводе, употреблении метафор и сопоставительном анализе метафорикона различных языков.

Толкования метафорических значений приводятся, как правило, с опорой на соответствующие толкования словарей литературного языка, однако нами была выполнена значительная работа по упорядочению толкований, устранению перекрестных отсылок и унификации толкований синонимичных метафор.

Перечень синонимов и эквивалентов метафор, приводимый в карточке, является уникальным разделом словаря, позволяющим не только переводить метафоры, но и исследовать метафорические модели, закономерности семантического варьирования, сопоставлять синонимические ряды метафор в белорусском и английском языках. На основе изучения области синонимов и эквивалентов можно обнаруживать пары слов с различным первичным и синонимическим (либо эквивалентным) метафорическим значением, например: *an* ***appetite*** *for life* и ***gravitation*** *towards computers* (значение ‘интерес’), *emotional* ***jail*** и *emotional* ***часцiт*** (значение ‘одиночество’), ***выбрык, фокус*** и ***вар’яцтва*** ‘фокус, сумасшествие’ (значение ‘не соответствующий нормам поступок’), *his proposals were too much* ***chaff*** и *адкінуць* ***шалупінне*** *нагавораў ‘отбросить шелуху клеветы’* (значение ‘бессмысленность, глупость’) (см. рис. 1).

Помимо ознакомления со значениями метафор и поиска синонимов и эквивалентов, пользователь имеет возможность сортировать данные в соответствии с источником и целью метафоризации. Для этого создана БД «Поиск», где путем выбора дескриптора можно осуществлять сортировку слов по определенному запросу.

Структура электронной карточки БД «Поиск» следующая: область дескрипторов, опция «Искать в найденном», стартовая кнопка «Поиск» и область для списка слов, удовлетворяющих заданным условиям.

В основе реестра дескрипторов, соотносимых с первичным (мотивирующим) значением слов, лежат четыре семантические сферы — «Природа», «Предмет», «Человек» и «Процессы и отношения». Внутри каждой сферы осуществляется тематическая градация на группы и подгруппы, например «Чело-

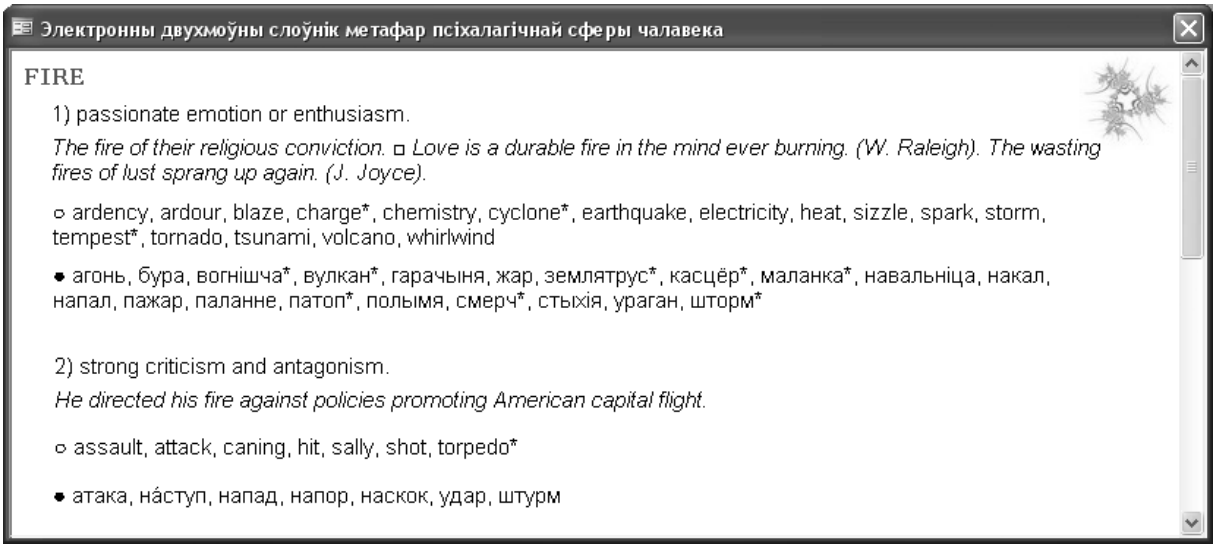


Рис. 1. Электронная карточка словаря

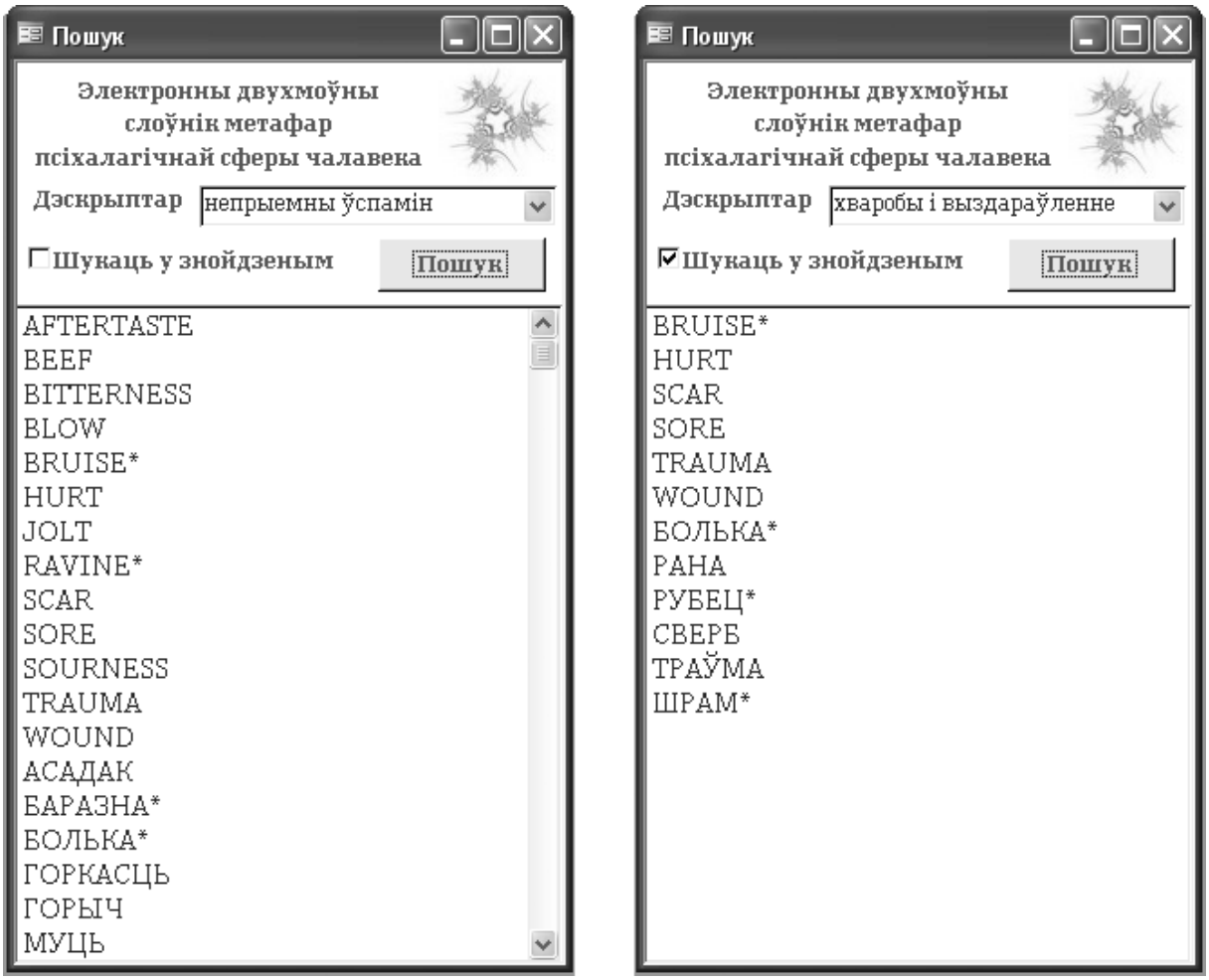


Рис. 2. Результаты комплексного поиска в БД «Поиск»

век — физиологические процессы и состояния — болезни и выздоровление» или «Природа — животный мир — группы животных и их жилище» и т. п.

Что касается структурирования сферы обозначаемого, то оказалось, что разграничить психический процесс и состояние либо психическое состояние и свойство личности часто невозможно (например, *непродолжительная* ***холодность*** *в отношениях* и ***холодность*** *— черта его характера*). Поэтому в реестре дескрипторов структура цели

метафорического переноса отражена в виде тематических групп (насколько это было возможно в связи с непредметным характером обозначаемого — см. об этом [5]): оценочные характеристики (соответствие / несоответствие нормам), наименования и характеристики психических процессов (мышление, память), психических состояний (радость, удовольствие, неудовлетворенность, страх, гнев и др.), свойств личности (характер, способности), отношений, поведения и речи. Однако в реестре есть дескрипторы, которые сложно отнести к какой-то одной из названных групп.

Невозможность строгой классификации целевой области приводит к некоторым трудностям на начальном этапе работы с БД «Поиск»: для осуществления выборки метафор пользователю необходимо ознакомиться с перечнем определяемых понятий и порядком их представления в словаре. Для большего удобства в реестр денотативных дескрипторов были включены варианты некоторых значений, например, «радость», «приятные чувства» и «удовлетворенность» (фактически один дескриптор), «плохое настроение» и «грусть» и некоторые другие.

БД «Поиск» позволяет осуществлять простой и комплексный поиск метафор с определенной семантикой. Например, выбор дескриптора «непрыемны ўспамін» ‘неприятное воспоминание’ позволит получить перечень метафор с соответствующим значением. Выбор опции «Искать в найденном» и дескриптора «хваробы і выздараўленне» ‘болезни и выздоровление’ поможет определить метафоры с соответствующим мотивирующим значением, используемые для номинации и характеризации неприятных воспоминаний (см. рис. 2).

Словарь позволяет выявлять случаи несовпадения переносных значений при совпадении первичных, например: *узнімаць са* ***дна*** *душы ўспаміны ‘поднимать со дна души воспоминания’ и his views have* ***bottom***. Обнаружение таких пар происходит в процессе ознакомления с перечнем эквивалентных метафор, в котором отсутствует ожидаемый пользователем вариант с эквивалентным первичным значением, что особенно важно при переводе слов с похожим звучанием. Так, например, метафора *mimicry* обозначает передразнивание в английском языке, а в белорусском языке *мімікрыя* — это приспособленчество и лицемерие: *a teasing* ***mimicry*** *in his words, гэтая пасада для яго —* ***мімікрыя*** *‘эта должность для него — мимикрия’.*

4. Методика создания словаря

В отличие от печатных словарей, где семантическое представление метафорической лексики может осуществляться довольно свободно — как путем отсылки к другим метафорам, так и путем декодирования метафоры, — в процессе создания электронного словаря метафор возникает вопрос об определении минимальной единицы, которая могла бы служить отправной точкой при систематизации первичного и непосредственно метафорического значения каждого полисеманта.

На первом этапе составления словаря мы опирались на дескрипторную теорию метафоры, разработанную А. Барановым [3], согласно которой каждой метафоре присваивался ряд денотативных (связанных с метафорическим значением) и сигнификативных (связанных с областью источника) дескрипторов. Так, например, метафора *dart* ‘a sudden, intense pang of a particular emotion’ имела денотативные дескрипторы <интенсивный>, <внезапный>, <эмоция> и сигнификативные <предмет>, <военные и охотничьи реалии>, <оружие и приспособления>. В процессе определения перечня денотативных дескрипторов мы выделили около 200 различных сем, таких как: ‘эмоция’, ‘внезапный’, ‘приятный’, ‘неприятный’, ‘непродолжительный’, ‘мышление’, ‘речь’, ‘несоответствие норме’ и др., некоторые из которых можно соотнести с описанными Ю.Д. Апресяном и А. Вежбицкой семантическими примитивами [4; 7]. В результате присвоения метафорам дескрипторов оказалось, что ввиду субъективности их определения и большого количества дескрипторов, приписываемых каждому отдельному метафорическому значению, упорядочение лексического материала не всегда соответствовало действительной картине семантических связей.

Невозможность упорядочить большое количество метафор с помощью метода дескрипторов вынудила нас искать иную минимальную единицу упорядочения значений, которой стало толкование. Разумеется, словарное толкование как оно есть не могло быть взято за абсолютную величину, поскольку большинство толкований метафор являются метафорическими и метод ступенчатой идентификации не позволяет выявить неметафорическое описание определенного понятия. В силу этого на основе обобщения словарных толкований нами были составлены около 100 базовых определений на белорусском языке, которые использовались в дальнейшем для упорядочения лексикографического материала, например (определения в статье даются в переводе на русский язык): «внезапное интенсивное проявление эмоций», «неприятное чувство как результат внешнего воздействия», «проявление неудовольствия во взгляде, выражении лица». Базовые толкования должны быть одинаковыми для слов с близкой семантикой, однако могут уточняться в каждой отдельной словарной статье, например: метафоры *nausea ‘a feeling of loathing and disgust’* и *allergy ‘a strong dislike’*, *алергія* ‘пачуццё агіды, непрыязнасці да каго- чаго-небудзь’ ‘аллергия — чувство отвращения, неприязни к кому- чему-л.’*

и *аскома ‘стамленне, незадаволенасць чым-небудзь’ ‘оскома — усталость, неудовлетворенность чем-л.’*, имеющие отличающиеся толкования в словаре, соотносятся с одним базовым толкованием — «отвращение, неприязнь».

Базовые определения основываются на среднем уровне абстракции используемых дескрипторов. С одной стороны, они не являются «тривиальными смыслами», перегружающими классификацию целевой области, а с другой — в отличие от толкований — они характеризуются большей схематичностью. Одно базовое определение используется для описания от 4 до 47 отдельных метафорических значений, например: «плохое настроение, грусть» (47 значений), «большое количество слов, впечатлений и т. п.» (41), «аффекты» (38), «внезапное неприятное чувство» (30), «проявление чувств во взгляде, выражении лица» (28), «противоречивые интенсивные эмоции» (21), «возникновение чувств, качеств» (19), «сентиментальность» (17), «активная позиция в споре» (16), «неожиданные поступки, противоречащие нормам» (12) и др. Что касается дескрипторов области источника, то в тематических группах обнаруживается от 4 (например, в группе «небесные тела») до 60 слов (например, в группе «вода и течение»).

В то время как словарь является результатом обработки многочисленного материала, он не может считаться исчерпывающим, поскольку выборка обусловлена составом словарей-источников, а обнаружение потенциальных языковых метафор, соответствующих регулярным переносам, может осуществляться бесконечно. Целевой поиск таких метафор осуществлялся в случае отсутствия как минимум одного узуально зафиксированного синонимического или эквивалентного значения, что обеспечило словарное отражение как минимум двух метафор с синонимическим значением в рамках одного языка. «Редкими» были базовые определения «передразнивание» (4 метафоры), «приспособленчество» (4), «отвращение» (4), «мечты» (6), «эмоциональная усталость» (6), «быстрая непрерывная речь» (6) и некоторые другие. Отметим, что номинативно автономные метафоры (например, *itch* в значении ‘желание’) имеют меньше синонимов и эквивалентов, а номинативно связанные (которые могут употребляться в сочетаниях с прямыми наименованиями явлений психологической сферы) — больше, например ***tornado*** *of confusion,* ***volcano*** *of resentment,* ***пажар*** *душы,* ***ураган*** *у думках ‘ураган в мыслях’*.

Как известно, компьютерная обработка данных предполагает установление соответствия между лингвистическим знаком (в данном случае, базовым толкованием) и математическим, поэтому каждому толкованию был присвоен определенный числовой код, использовавшийся при дальнейшей систематизации метафор в базе данных. Помимо кода толкования каждому слову присваивался код языка («белорусский» или «английский»), а также код значения. Подбор синонимов одного языка и эквивалентов другого осуществляется Запросами: при совпадении кода толкования слова́ с совпадающим кодом языка определяются как синонимы, а несовпадающим — как эквиваленты.

Тестирование словаря показало, что метод толкований является достаточно эффективным для составления подобного электронного ресурса: соответствие нескольких метафор одному общему значению и их появление в перечне синонимов и эквивалентов минимализирует процент «лишних» слов и, кроме того, способствует упорядочению самих толкований в словаре, на необходимость чего указывают многие лексикографы. В то же время следует отметить, что следующий за обнаружением синонимических (или эквивалентных) пар метафор этап предполагает обращение к контексту — сочетаемости метафор в текстах, на что ориентирован иллюстративный материал словаря, который планируется значительно расширить.

5. Заключение

Новый ресурс — Электронный двуязычный словарь метафор психологической сферы человека — позволяет наиболее полно отразить средства метафорической репрезентации внутреннего мира человека в белорусском и английском языках, выявить синонимические и эквивалентные отношения метафор, а также осуществить сортировку и поиск метафор в соответствии с запросом пользователя. Разработанная методика упорядочения и сортировки метафорических значений с опорой на базовые толкования может быть использована для обработки различных видов лексикографической информации.

Литература

1. *Metaphors* Dictionary. Ed. by E. Sommer, D. Weiss // Detroit: Visible Ink Press, 1994.
2. *Online* Dictionary, Encyclopedia and Thesaurus. Free Access // http://www.thefreedictionary.com.
3. *Oxford* Dictionary of English. Ed. by C. Soanes, A. Stevenson // Oxford: Oxford University Press, 2003.
4. *Апресян В. Ю.* Метафора в семантическом представлении эмоций // Вопросы языкознания. 1993. № 3. С. 27–35.
5. *Апресян Ю. Д.* О языке толкований и семантических примитивах // Известия академии наук. Серия литературы и языка. 1994. № 4. С. 27–41.
6. *Баранов А. Н.* Дескрипторная теория метафоры и типология метафорических моделей // http://www.uisrussia.msu.ru/linguist/_A1_2_4_1_metaphor.jsp.
7. *Вежбицкая А.* Семантические универсалии и описание языков // М.: Языки славянской культуры, 1999.
8. *Слоўнік* мовы Янкі Купалы: у 8 т. Рэдкал.: У. В. Анічэнка [і інш.] // Мінск: Беларуская навука, 1997–2002.
9. *Старычонак В. Д.* Слоўнік метафар беларускай мовы [рукапіс] // Мінск, 2001.
10. *Тлумачальны* слоўнік беларускай літаратурнай мовы. Пад рэд. М. Р. Судніка, М. Н. Крыўко. // Мінск: БелЭн, 2002.
11. *Тлумачальны* слоўнік беларускай мовы: у 5 т. Пад аг. рэд. К. К. Атраховіча. // Мінск: БелСЭ, 1977–1984.

Видовая корреляция в толковом словаре*

Aspectual correlation in an explanatory dictionary

Шмелев А. Д. (shmelev.alexei@gmail.com)

Московский педагогический государственный университет;
Институт русского языка им. В. В. Виноградова РАН, Москва

Рассматривается распространенная лексикографическая практика, в соответствии с которой полное толкование дается только одному из членов видовой пары, тогда как для другого вместо толкования дается отсылка. Обсуждается вопрос о том, в каких случаях такая практика допустима или даже необходима.

1. Природа русского глагольного вида

Как известно, русский глагольный вид одни грамматисты считают словоизменительной, а другие — классифицирующей категорией. В первом случае говорят, что одна и та же глагольная лексема может иметь формы как совершенного, так и несовершенного вида; во втором случае говорят о глагольных лексемах совершенного и несовершенного вида, которые могут быть связаны между собою какими-то (напр., словообразовательными) отношениями. Сама проблема связана с фундаментальным свойством русских глаголов, состоящим в том, что любая глагольная форма, употребленная в высказывании, относится к тому или иному виду и при этом абстракция отождествления глагольных форм одного вида в качестве представителей одной лексемы не вызывает затруднений. Совокупность глагольных форм одного вида, отождествленных в качестве представителей одной лексемы, назовем глаголом совершенного и, соответственно, несовершенного вида соответственно. Значительно менее ясным является вопрос, возможно ли (и если да, то в каких случаях) отождествление разновидовых глаголов в качестве представителей одной лексемы. Признание того, что это возможно и означает принятие словоизменительной трактовки глагольного вида. При таком подходе говорят также, что разновидовые глаголы, объединяемые в одну лексему составляют видовую пару, а отношение между этими глаголами можно назвать отношением видовой коррелятивности. Понятие видовой пары не менее активно используется и при принятии трактовки глагольного вида как классифицирующей категории. В этом случае предполагается, что между разновидовыми глаголами может иметь место особо тесная связь (видовая коррелятивность), которая позволяет попарно объединить эти глаголы, но уже не в одну лексему, а в видовую пару[1].

В книге [Зализняк, Шмелев 2000] мы высказали соображение, что все существенное, что можно сказать о русском виде вообще и видовой коррелятивности в частности, может быть сформулировано в терминах и той, и другой трактовки вида (как словоизменительной или классифицирующей категории). Выбор между ними определяется не столько сутью дела, сколько удобством описания. В частности, в толковом словаре достаточно дать толкование одному из членов видовой пары, если из этого толкования можно вывести толкование другого члена той же пары при помощи регулярных и универсальных

---

* Настоящая статья развивает положения, сформулированные в ряде предшествующих публикаций, в частности [Шмелев 1989; Булыгина, Шмелев 1992; Зализняк, Шмелев 2000, 97–103; Mikaelian, etc. 2008; Шмелев 2008аб].

[1] Заметим, что при принятии трактовки глагольного вида как классифицирующей категории встречается подход, не признающий существования видовых пар и, соответственно, видовой коррелятивности. Сторонники такого подхода объявляют понятие видовой пары устаревшим и предлагают отказаться от него в пользу понятия «видового кластера» [Janda 2007]. Заметим, что сам термин «видовой кластер» представляется не слишком удачным. Что в нем значит прилагательное «видовой»? Когда говорят о видовой паре или видовой коррелятивности, слово «видовой» указывает на то, что члены пары или корреляты противопоставлены по виду; но члены «кластера» не могут быть противопоставлены по виду, поскольку в русском языке только два вида.

правил. В этом случае даже при морфологической нерегулярности в словарной статье нетолкуемого члена можно ограничиться отсылкой к толкованию его видового коррелята. Такое решение будет практически эквивалентно признанию словоизменительной природы глагольного вида. Напротив того, невозможность вывести при помощи регулярных и универсальных правил толкование одного видового коррелята из толкования другого означает необходимость давать в толковом словаре и то, и другое толкование. В этом случае принятие словоизменительной трактовки вида будет предполагать соединение в рамках одной лексемы двух несводимых друг к другу толкований, при этом различие толкований соответствует формальному видовому различию. Представляется, что в этом случае значительно удобнее иметь две независимых словарных статьи, что практически эквивалентно принятию трактовки вида как классифицирующей категории.

2. Отсылки или полные толкования?

Независимо от того, считать ли глагольный вид словоизменительной или классифицирующей грамматической категорией, все, кто всерьез изучал глагольную семантику, даже если признают, существование видовых пар, знают, что даже между членами видовых пар самых регулярных семантических типов, как правило, имеются важные смысловые различия. Дело в том, что глаголы совершенного вида обозначают события, т. е. предполагают, что ранее имело место одно положение дел, а после того, как произошло данное событие, возникло другое, новое положение дел. В то же время для глаголов несовершенного вида, в том числе входящих в видовые пары, характерны употребления иного типа, соотносимые не с событиями, а с явлениями, длящимися в определенный момент или промежуток времени (процессами и состояниями), а также с вневременными свойствами. Говорить ли при этом о семантическом различии форм совершенного и несовершенного вида одного и того же глагола или о семантическом различии двух коррелятивных глаголов совершенного и несовершенного вида, считать ли эти различия различием в лексическом или грамматическом значении, — вопрос отдельный и не отменяющий самого факта существования различий. При этом сформулировать универсальные правила, которые позволяли бы выводить толкование глагола одного вида из толкования глагола противоположного вида, оказывается затруднительно.

Рассмотрим семантические различия между глаголами *решить* и *решать <задачу>*, описанные Ю. Д. Апресяном при помощи следующих двух толкований: «*X решает P* ≈ ‘X обдумывает информацию, имеющую отношение к P, с целью получить ответ на содержащийся в P вопрос’; *X решил P* ≈ ‘X решал P [пресуппозиция]; X получил ответ на содержащийся в P вопрос [ассерция]’» [Апресян 2005, 41–42].

Можно заметить, что использование отсылки ‘сов. к *решать <P>*’ вместо полного толкования глагола *решить <P>* могло бы быть допустимо только в том случае, если бы мы располагали правилами, посредством которых, имея эту отсылку и толкование глагола *решать <P>* (напр., ‘обдумывать информацию, имеющую отношение к P, с целью получить ответ на содержащийся в P вопрос’), мы могли бы получить толкование глагола *решить <P>*. Это могло бы быть, если бы мы сформулировали напр., такое правило: «если глагол несовершенного вида содержит указание на цель, с которой производится действие, то у его перфективного коррелята указание на действие становится пресуппозицией, а ассерцией становится указание на достижение цели». Однако попытка использовать такое правило связана со значительными трудностями, поскольку множество имперфективных глаголов подразумевают целенаправленное действие, однако их перфективные корреляты не предполагают достижения цели. Чтобы не ходить далеко за примерами, рассмотрим толкование глагола *жаловаться* в уже цитированной статье Ю. Д. Апресяна: «Лексема *жаловаться 1* значит ‘говорить, что произошло или имеет место нечто плохое для субъекта, чтобы побудить адресата речи исправить положение или найти у него понимание’» [Апресян 2005, 38]. Мы видим, что глагол *жаловаться* в рассматриваемом значении содержит ясное указание на цель. Исходя из этого, мы могли бы предположить, что если сформулированное выше правило верно, то его перфективный коррелят *пожаловаться* должен указывать на достижение цели и значить нечто вроде ‘говоря, что произошло или имеет место нечто плохое для субъекта, побудить адресата речи исправить положение или найти у него понимание’. Однако глагол *пожаловаться* такого значения не имеет: он указывает на речевое действие, произведенное с некоторой целью, но ничего не сообщает о том, была ли цель достигнута[2].

Использование отсылки ‘несов. к *решить <P>*’ вместо полного толкования глагола *решать <P>* наталкивается на сходные трудности. Оно могло бы быть допустимо, если бы мы сформулировали напр., такое правило: «если глагол совершенного вида указывает на достижение результата посредством некоторого действия, то у его имперфективного коррелята указание на действие из пресуппозиции становится ассерцией, а достигнутый результат превращается в цель, с которой производится

[2] С другой стороны, как известно, пары разновидовых глаголов, в которых глагол несовершенного вида указывает на деятельность, направленную на достижение некоторой цели, а глагол совершенного вида на достижение этой цели, вовсе не обязательно являются видовыми парами.

действие». Однако для множества глаголов такое «правило» даст неверный результат. Скажем глагол, *выиграть у Y-а в шахматы* значит нечто вроде ‘играя с Y-ом в шахматы, победить Y-а’; однако *выигрывать у Y-а в шахматы* вовсе не значит ‘играть с Y-ом в шахматы с целью победить Y-а’.

Могло бы показаться, что отказ от отсылок и использование полных толкований для глаголов обоих видов решит все трудности, и это могло бы быть серьезным аргументом против словоизменительной трактовки вида и самого понятия видовой коррелятивности. Однако если мы дадим независимые толкования обоим членам видовой пары без указания на видовую коррелятивность, толкования во многих случаях будут не соответствовать реальному языковому употреблению соответствующих глаголов. Так, если мы истолкуем глагол *решать <P>* как ‘обдумывать информацию, имеющую отношение к P, с целью получить ответ на содержащийся в P вопрос’, а глагол *решить <P>* как ‘обдумывая информацию, имеющую отношение к P, получить ответ на содержащийся в P вопрос’, то окажется, что толкования не соответствую его употреблению, напр., в таких предложениях:

(50) *Он с ходу решает второе уравнение и переходит к третьему.* ≈ ‘Он… получает ответ на содержащийся во втором уравнении вопрос…’

(51) *Как только студенты вспоминали эту теорему, они легко решали задачу.* ≈ ‘…Они получали ответ на содержащийся в задаче вопрос…’

Следующая пара предложений почти синонимична, хотя в первом из них употреблен глагол совершенного вида, а во втором — несовершенного (причем речь в них идет именно о получении ответа на содержащийся в задаче вопрос, а не об «обдумывании»):

(52) *Стоило студенту решить хотя бы одну задачу, преподаватель ставил ему зачет.*

(53) *Как только студент решал хотя бы одну задачу, преподаватель ставил ему зачет.*

Чтобы учесть эти употребления, необходимо предусмотреть в описании глагола несовершенного вида возможность его использования в значении, соответствующем толкованию его перфективного коррелята; тем самым нельзя отказаться от упоминания видовой коррелятивности в словарной статье.

3. Обязательная имперфективация

Дело в том, что необходимость понятия видовой корреляции вытекает из самого устройства русской аспектуальной системы. При описании русской грамматики исследователь стоит перед необходимостью моделировать явление, которое в совместных публикациях с Анной А. Зализняк мы называли «обязательной имперфективацией». Речь идет о том, что в определенных случаях (в частности, в настоящем времени, при обозначении неоднократных или узуальных действий или событий и др.) правила русской грамматики предписывают использование несовершенного вида. В таких случаях, даже если «в норме» нужный смысл выражается глаголом совершенного вида, говорящий обязан найти этому глаголу имперфективное соответствие. При такой замене семантические отличия глагола несовершенного вида от его перфективного коррелята тривиальны: вносится значение неоднократности, изобразительности и т. д., а само обозначаемое событие остается тем же самым. Это и приводит к тому, что носители языка часто вообще не замечают семантического сдвига и не отдают себе отчет в том, что меняется вид. Понятие видового коррелята (и, соответственно, видовой пары) как раз и представляет собою инструмент описания указанной способности носителя языка: мы говорим, что регулярной заменой для глагола совершенного вида является его имперфективный коррелят (с которым он образует видовую пару). Критерий видовой парности, сформулированный в знаменитой статье [Маслов 1948], непосредственно вытекает из определения видовых пар. Необходимо подчеркнуть, что этот критерий не претендует на то, чтобы быть операционным; его не следует воспринимать как инструмент, используемый для выявления видовых пар, которые нам необходимы для какой-то таинственной и непонятной цели, — он просто следует из определения видовых пар, которые как раз и являются инструментом описания «обязательной имперфективации».

Подчеркнем, что, говоря об обязательной имперфективации, мы понимали термин «имперфективация» строго функционально — как «замену» глагола совершенного вида на его видовой коррелят. Наряду с этим термин «имперфективация» используется и по отношению к морфологическому явлению — образованию глагола несовершенного вида (вторичного имперфектива) от глагола совершенного вида при помощи имперфективирующего суффикса (в этом смысле можно говорить о морфологической имперфективации).

В большинстве случаев, если от глагола совершенного вида можно образовать имперфектив посредством морфологической имперфективации, то именно этот имперфективный глагол и является его коррелятом (т. е. используеся при обязательной имперфективации). Исключения немногочисленны и хорошо известны: в качестве имперфективных коррелятов производящих глаголов совершенного вида не используются такие вторичные имперфек-

тивы, как *умолять*, *заблуждаться*, *сказывать*, *взимать*[3].

Однако существует и другой механизм имперфективации, а именно — депрефиксация. Сущность ее состоит в том, что в качестве имперфективного коррелята приставочного глагола совершенного вида используется соответствующий бесприставочный глагол (несовершенного вида). В результате действия механизма депрефиксации мы получаем такие видовые пары, как сделать/делать, сыграть/играть, увидеть/видеть, испугаться/пугаться, потерять/терять, поменять/менять, попросить/просить, построить/строить, пожаловаться/жаловаться (особенно характерен этот механизм для глаголов с приставкой *по-*). Следует иметь в виду, что депрефиксация представляет собою именно функциональный механизм, посредством которого мы получаем имперфективный коррелят приставочного перфективного глагола при обязательной имперфективации. С морфологической точки зрения дело обстоит противоположным образом: от бесприставочного глагола несовершенного вида образуется глагол совершенного вида; иными словами, имеет место префиксация.

В некоторых случаях, когда в нашем распоряжении нет имперфективного коррелята, полученного посредством морфологической имперфективации или посредством депрефиксации, используется супплетивный коррелят — отсюда такие видовые пары, как *поймать/ловить*, *положить/класть*, *взять/брать*, *сказать/говорить*.

Выбор механизма имперфективации может зависеть от субъективных оценок говорящего. Так, для глагола *похерить* одни носители языка предпочитают имперфективный коррелят *похеривать*, а другие — *херить*:

(54) *Сознаюсь, я доводил эту фантазию до таких окраин, что похеривал даже самое образование.*
(Достоевский)

(55) *Когда-то, правда, был один — Ему бы отдалась я, может, — А прочих херю я мужчин: Дотронуться никто не может.*
Евгений Кропивницкий

Можно сказать, что единственным (но неустранимым) основанием для включения в описание указания на коррелятивность разновидовых глаголов является необходимость описания обязательной имперфективации. Для описания прочих языковых явлений, связанных с видом, существенно более эффективно опираться на толкования глаголов обоих видов. Тем самым критерий Маслова следует рассматривать не просто как удобный способ установить видовую коррелятивность, но как описание самого ее существа.

4. Тривиальные и нетривиальные значения

При обязательной имперфективации семантические отличия глагола несовершенного вида от его перфективного коррелята полностью предсказуемы и регулярны; поэтому их и называют «тривиальными». В тривиальных значениях имперфективный глагол является своего рода «двойником» своего перфективного коррелята, почти дублируя не только его семантику, но и синтаксические свойства.

Однако для имперфективных членов большинства видовых пар характерно наличие «нетривиальных» значений, которые могут существенным образом отличаться от своих перфективных коррелятов как в отношении семантики, так и прочих языковых свойств. Эти отличия гораздо в большей степени бросаются в глаза, и нередко, когда говорят о семантических различиях соотнесенных глаголов совершенного и несовершенного вида, подразумевается (хотя эксплицитно не оговаривается), что рассматриваются «нетривиальные» различия, а случаи, когда имеет место обязательная имперфективация и различия «тривиальны», в расчет не принимаются. Легко видеть, что приведенное выше толкование Ю. Д. Апресяна для глагола *решать* (в форме несовершенного вида), имеющее «нетривиальные» отличия от толкования глагола *решить*, не относится к употреблениям глагола *решать* как раз в тех случаях, когда выбор несовершенного вида обусловлен обязательной имперфективацией (примеры (1), (2) и (4)).

Можно добавить, что синтаксические свойства имперфективных глаголов в тривиальных употреблениях в целом совпадают со свойствами их перфективных коррелятов (Булыгина, Шмелев 1992). Например, глагол *решить* в значении ‘обдумывая вопрос Q, прийти к выводу P’ может подчинять косвенный вопрос (*решили, кто пойдет за водкой*), изъяснительное придаточное (*решили, что за водкой пойдет Петя*) и инфинитив (*решили послать за водкой Петю*); точно такими же синтаксическими возможностями обладает его имперфективный коррелят *решать* в тривиальном употреблении (напр., в настоящем историческом: *После долгих обсуждений они наконец решают, кто пойдет за водкой*; *После долгих обсуждений они наконец решают, что за водкой пойдет Петя*; *После долгих обсуждений они наконец решают послать за водкой Петю*). В нетривиальных значениях грамматические свойства глагола несовершенного вида могут суще-

[3] При обязательной имперфективации в качестве коррелятов глаголов умолить и заблудиться иногда используются окказиональные вторичные имперфективы умаливать и заблуживаться, а глаголы сказать и взять имеют регулярные супплетивные корреляты говорить и брать.

ственно отличаться от свойств его перфективного коррелята. Так, глагол *решать* в нетривиальном актуально-длительном значении ‘обдумывать вопрос Q с целью прийти к какому-либо выводу’ может подчинять косвенный вопрос (*Они сейчас решают, кто пойдет за водкой*), но не изъяснительное придаточное (**Они сейчас решают, что за водкой пойдет Петя*) и не инфинитив (**Они сейчас решают послать за водкой Петю*).

Для целого ряда видовых пар семантические и синтаксические особенности имперфективных членов в нетривиальных значениях оказываются идиосинкратическими и очень сильно отличаются от семантических и синтаксических свойств их перфективных коррелятов. Проиллюстрируем это на примере видовой пары *решиться/решаться*[4], *которая обладает совсем иными особенностями* по сравнению с внешне сходной парой *решить/решать*. *Решаться* (без отрицания) выступает преимущественно как «двойник» своего перфективного коррелята (напр., при обозначении повторяющегося события). Ср.: … *они мгновенно подчиняют податливых матерей своей железной воле и перебираются из жесткой кровати на мягкие руки. Тогда матери приходится надолго забыть о кровати; она спит в кресле, под ноющую спину подложена подушка, на груди расположился любимый тиран, пошевелиться нельзя. Некоторые добиваются еще большего, заставляют носить их на руках ночь напролет; как только мама присядет, начинается нестерпимый вой. Но это уже совсем беспредел, решаются на него, как правило, только мальчики* (Александр Архангельский). Этот глагол не склонен употребляться в актуально-длительном значении (сомнительно:*?Не мешай: они там заняты важным делом — решаются присудить премию неизвестному молодому ученому*). Однако он может обозначать нерешимость субъекта, преодоление которого является условием наступления события, обозначаемого перфективным членом (*долго решался и наконец решился*) — ср.: …*бабки долго решались и вдруг, как сговорились, скопом запросили: «Благослови, отрок!»* (Петр Алешковский). Чаще всего форма несовершенного вида употребляется с отрицанием и при этом также обозначает концептуально выделенное состояние нерешимости (*Ну как, вы все не решаетесь принять его предложение?*); таким образом возникает парадоксальная энантиосемия, когда один и тот же глагол с отрицанием и без отрицания значит почти одно и то же: *долго решался войти ≈ долго не решался войти*. Ср.: *Я долго не решалась писать к тебе, — в продолжение двух месяцев каждый день я бралась за перо и бросала его…* [*≈ долго решалась написать к тебе…*] (Панаев); *Не мало времени, не мало убеждений и просьб стоило Марье Ивановне, чтоб уговорить Дуню идти в столовую и познакомиться с Денисовым. Долго не решалась Дуня, наконец пересилила себя — пошла* [*≈ долго решалась, наконец пересилила себя — пошла*] (Андрей Печерский)[5].

5. Лексикографические и теоретические следствия

Отсылочную помету «несов. к…» целесообразно использовать для того, чтобы указать на тривиальные значения глагола несовершенного вида, соответствующие значению перфективного коррелята, к которому производится отсылка. Действительно, в этом случае толкование глагола несовершенного вида получается из толкования перфективного коррелята посредством простых, «тривиальных» операций. Синтаксические свойства парного глагола несовершенного вида в тривиальных значениях также «дублируют» синтаксические свойства его перфективного коррелята. Напротив того, толкования нетривиальных значений парных глаголов несовершенного вида вывести из толкований их перфективных коррелятов значительно труднее, или вовсе невозможно (как для глагола *решаться*); поэтому целесообразно для этих значений приводить полное толкование[6].

Для словарей активного типа необходимо также указание на имперфективный коррелят, поскольку знание имперфективного коррелята данного глагола необходимо для «автоматической замены» в позиции обязательной имперфективации. Тогда для глагола *решить <вопрос Q>* толкование (приблизительно ‘обдумав вопрос Q, прийти к определенному выводу’) должно сопровождаться указанием: «несов. *решать*».

Итак, толкование глагола *решать <вопрос Q>* может выглядеть как ‘1. несов. к *решить <вопрос Q>*; 2. обдумывать вопрос Q с целью прийти к какому-либо выводу’; толкование глагола *решать <задачу P>* — как ‘1. несов. к *решить <задачу P>*; 2. обдумывать информацию, имеющую отношение к P, с целью

---

[4] Эта пара рассматривалась нами в статье (Булыгина, Шмелев 1992); здесь в предложенное описание вносится ряд уточнений.

[5] Впрочем, определенное семантическое различие между формами без отрицания и с отрицанием все же наблюдается. Выражение долго решался в большей степени предполагается некоторый внутренний процесс, когда субъект собирается с духом, прежде чем предпринять решительное действие, тогда как долго не решался скорее указывает на состояние, когда субъект еще никак не готов к действию и, скорее всего, еще не начал процесс внутренней подготовки к действию.

[6] Помимо этого, если в словаре предполагается описывать синтаксические свойства толкуемых глаголов, то синтаксические свойства парных глаголов несовершенного вида в нетривиальных значениях также следует описать отдельно.

получить ответ на содержащийся в P вопрос'. (Разумеется, конкретные детали толкования «нетривиального значения зависят от типа и предназначения словаря.) Разные «значения несовершенного вида» тем самым рапсределяются по разным лексическим значениям. И лишь для парных глаголов несовершенного вида, которые способны только к тривиальным употреблениям, можно ограничиться отсылкой к перфективному корреляту: *находить* 'несов. к *найти*'; *приходить* 'несов. к *прийти*'; *прибегать* 'несов. к *прибежать*'; *добегать* 'несов. к *добежать*'.

Собственно, сходные лексикографические решения предлагались и в предшествующих работах, упомянутых в сноске к заглавию данной статьи. Однако там отсылку к перфективному корреляту и «полноценное толкование глагола несовершенного вида предлагалось подавать в рамках одного лексического значения (и тем самым предполагалось, что «тривиальные» и «нетривиальные» значения несовершенного вида могут принадлежать одному лексическому значению). Здесь я готов пойти дальше и, исходя из того, что различия между что «тривиальными» и «нетривиальными» значениями столь разнообразны и непредсказуемы, что описание их в рамках одного лексического значения либо оказывается чрезмерно громоздким, либо приводит к потере релевантной информации.

Можно сформулировать и теоретический вывод. Двойственность трактовки категории глагольного вида как словоизменительной и классифицирующей отражает реальную двойственность ее функционирования в языке. В тех случаях когда видовое противопоставление используется для того, чтобы осуществить замену вида в позиции «обязательной имперфективации», мы имеем дело с видовой коррелятивностью и вид ведет себя как словоизменительная категория (и ощущается таковой многими «наивными» носителями языка). В тех же случаях когда противопоставление по виду используется для того, чтобы отразить онтологические различия обозначаемых явлений, вид ведет себя как классифицирующая категория: разновидовые глаголы, даже если мы их объединили в видовую пару на основании критерия Маслова имеют отчетливо разное значение, должны получить разные толкования, не сводимые друг к другу, и независимое лексикографическое описание. Фактически в этом случае мы имеем дело с разными лексическими значениями глагола.

Литература

1. *Апресян Ю. Д.* Правила взаимодействия значений и словарь // Русский язык в научном освещении. 2005, № 1(9). С. 7–45.
2. *Булыгина Т. В., Шмелев А. Д.* Идентификация событий: онтология, аспектология, лексикография // Логический анализ языка: Модели действия, М.: Наука, 1992. С. 108–115.
3. *Зализняк Анна А., Шмелев А. Д.* Введение в русскую аспектологию // М.: Языки русской культуры, 2000.
4. *Маслов Ю. С.* Вид и лексическое значение глагола в современном русском литературном языке // Известия АН СССР. Серия литературы и языка. 1948. Т. 7, № 4. С. 303–316.
5. *Шмелев А. Д.* Видовые пары в базовом толковом словаре // Русский язык в условиях двуязычия и многоязычия: Проблемы функционирования и исследования: Тезисы докладов VI Всесоюзного совещания. Минск: Навука і тэхніка, 1990. С. 94–95.
6. *Шмелев А. Д.* Имперфективизация и видовая корреляция // Aspekte, Kategorien und Kontakte slavischer Sprachen. Festschrift für Volkmar Lehmann zum 65. Geburtstag. Hamburg: Verlag Dr. Kovač, 2008a. S. 372–379.
7. *Шмелев А. Д.* Имперфективация и толковый словарь // С любовью к слову. Festschrift in Honour of Professor Arto Mustajoki on the Occasion of his 60th Birthday. Helsinki: Department of Slavonic and Baltic Languages and Literatures, 2008б. P. 362–370.
8. *Mikaelian I., Shmelev A., Zalizniak Anna A.* Le concept de couple aspectuel, est-il encore utile? // Questions de linguistique slave. Etudes offertes à Marguerite Guiraud-Weber. Aix-en-Provence, 2008. P. 189–206.

Письменное бытование русского анекдота*

Russian jokes in written form

Шмелева Е. Я. (eshkind@mail.ru),
Шмелев А. Д. (shmelev.alexei@gmail.com)
Институт русского языка им. В. В. Виноградова РАН, Москва

Обсуждается письменное бытование русских анекдотов. Предлагается различать фиксацию рассказанного устно анекдота на письме; «протоанекдоты»; анекдоты, бытующие в Интернете.

1. Вступительные замечания

Рассказывание анекдота представляет собою исключительно устный речевой жанр. Мы неоднократно настаивали на этом в целом ряде публикаций и обсуждали вытекающие из этого особенности анекдота как текста и как речевого жанра. В частности, мы отмечали, что по своим языковым особенностям анекдот относится не к повествовательным, а, скорее, к «изобразительным» жанрам. Рассказывание анекдота — это не повествование, а своего рода представление, производимое единственным актером. Хорошо рассказать анекдот — значит не просто осуществить повествование о некотором забавном эпизоде, но представить этот эпизод «в лицах». Кроме того, для многих анекдотов первостепенную конструктивную роль играет интонация рассказчика, его мимика и жестикуляция, и в ряде случаев именно они создают то, что называется «солью» анекдота [Шмелева, Шмелев 1998: 264; 2002: 24–25] (жестовой составляющей анекдотов и ее взаимодействию со словесным оформлением была специально посвящена статья [Крейдлин, Шмелева 2007]).

В то же время в последние два десятилетия чрезвычайно широкое распространение получило функционирование анекдотов в качестве записанных текстов. Показательно, что сочетание *прочесть анекдот*, которое казалось странным в шестидесятые-семидесятые годы XX в.[1], стало весьма частотным в современных текстах. При этом некоторое отклонение от традиционной нормы в сочетании *прочесть анекдот*, продолжает ощущаться многими носителями языка; иногда это ощущение проявляется в утверждениях, что *прочтенные* анекдоты редко бывают остроумными и, как правило, не могут вызвать смеха. Ср. высказывание Владимира Войновича в интервью, данном газете «Аргументы и факты» в 2001 г. На замечание интервьюера (Ольги Шигаревой): «Меня всегда поражало, почему люди с такими серьезными лицами читают анекдоты», — Войнович отвечает: «Это вообще загадка анекдота — в написанном виде он “не работает”. Не знаю, чем это объяснить, но написанные на бумаге анекдоты редко смешат. Лучшее время для анекдота — глухое». Любопытно, что сказанное касается и тех, анекдотов, которые, будучи рассказанными устно, кажутся слушателям и остроумными, и смешными. Примечательно, что «метатекстовый ввод», который, как мы отмечали [Шмелева, Шмелев 1998: 264; 2002: 29–31], всегда предваряет рассказывание анекдота, может включать такие вопросы, как «Слыхал анекдот?», но практически никогда не включает вопрос «Читал анекдот?» Тем не менее сама идея чтения анекдотов уже не является в настоящее время чем-то из ряда вон выходящим, и не случайно она отражается даже в названиях некоторых сайтов в Интернете (напр., http://anekdot2.com/prochitajte-eti-anekdoty).

Эта новая форма бытования анекдотов не осталась незамеченной специалистами. В статье, опубликованной в журнале «Новое литературное обозрение» в 1996 г., А. В. Вознесенский заговорил о новом явлении, «анекдотопечатании», возникшем в 1990 г., на волне «перестройки»; автор сообщал, что всего за шесть лет (1990–1995) собрал в своей коллекции 550 печатных сборников анекдотов (без учета периодики) [Вознесенский 1996: 396].

---

* Данная статья продолжает серию докладов на конференции «Диалог», начатую в 1998 г. (см. [Шмелева, Шмелев 1998]).

[1] Оно вполне могло встретиться в текстах XIX и начала XX вв., но в этом случае слово *анекдот* понималось в «старом смысле» — как занимательное описание действительного события. Ср.: *Еще недавно я читал один анекдот, которому бы не поверил, если б не знал, что он совершенная правда* (Федор Достоевский)

Но, разумеется, письменное бытование анекдотов не ограничивается «анекдотопечатанием». Еще более существенным следует признать обилие сайтов Интернета, на которых анекдоты представлены именно в письменной форме.

В настоящей заметке мы рассмотрим основные типы бытования анекдотов на письме и некоторые языковые особенности письменного текста анекдотов на фоне языковой специфики анекдота как устного речевого жанра.

2. Письменная фиксация анекдота

Как и любые произведения устных речевых жанров, рассказанные анекдоты допускают письменную фиксацию. Такая фиксация может осуществляться как с научными, так и с развлекательными целями.

Следует различать два типа письменной фиксации анекдота: с одной стороны, может фиксироваться конкретный акт рассказывания анекдота, а с другой — может записываться инвариантный текст анекдота, реализующийся в конкретных актах рассказывания. При фиксации конкретного акта рассказывания могут указываться разнообразные параметры данного акта (время и место рассказывания, ситуация, анкетные данные рассказчика, использованный «метатекстовый ввод», жесты, которыми рассказчик сопровождал анекдот, слова-паразиты, вставляемые им в речь, интонация, паузы хезитации, автокоррекция, если рассказчик в какой-то момент сбился, реакция слушателей и т. д.). В зависимости от целей такой фиксации те или иные параметры могут опускаться: так, для социолога могут быть безразличны слова-паразиты, интонация или паузы хезитации, представляющие первостепенный интерес для лингвиста. Для большинства целей оказывается возможным отвлечься от особенностей конкретного акта рассказывания и ограничиться фиксацией инвариантного текста анекдота, взятого в отвлечении от отдельных актов его рассказывания. В этом случае не описывается ситуация рассказывания анекдота, не указываются время и место рассказывания и анкетные данные рассказчика. «Метатекстовый ввод», жесты, интонация приводятся лишь в тех случаях, когда они оказываются ингерентной составляющей инвариантного текста анекдота (разнообразные примеры такого рода были приведены в книге [Шмелева, Шмелев 2002: 24–25, 29–31]). При этом указание на жесты или интонацию осуществляются посредством специальных помет (ремарок), которые тем самым могут рассматриваться как дополнительная, четвертая составляющая текста анекдота (напомним, что для устного рассказывания анекдота мы выделяли три составляющих: «метатекстовый ввод», текст «от повествователя», речь персонажей). Если при фиксации конкретного акта рассказывания описание параметров данного акта делается как бы от лица наблюдателя, то при фиксации инвариантного текста ремарки служат скорее инструкцией рассказчикам.

Заметим, что фиксация конкретного акта рассказывания чаще всего осуществляется с научными целями; вне таких целей она используется редко (напр., в рамках репортажа или дневниковой записи) и, как правило, далека от точности. Обычно «наивная» фиксация анекдота ограничивается текстом анекдота, отвлеченным от конкретных актов рассказывания, т. е. приближается к инвариантному тексту анекдота. Именно такие тексты чаще всего публикуются в разнообразных сборниках анекдотов, т. е. в рамках «анекдотопечатания»[2]. Отметим, что ремарки в таком случае обычно встраиваются в собственный текст анекдота и внешне не отделяются от текста повествователя. Ср. пример из книги [Ничипорович1998: 15]:

(56) *Новый русский приходит в роддом, его встречает акушерка. — Поздравляю, у вас мальчик! — И сколько? — Три пятьсот! — Не вопрос! — произносит новоиспеченный папаша, доставая кошелек.*

Понятно, что при рассказывании этого анекдота рассказчик не будет говорить слова *произносит новоиспеченный папаша, доставая кошелек*; вместо этого он должен жестами имитировать доставание кошелька и отсчитывание денег.

Следует иметь в виду, что в сборниках анекдотов, наряду с инвариантными текстами анекдотов, часто приводятся тексты, которые без более или менее существенной модификации не могут быть рассказаны как анекдот (в интересующем нас смысле слова). Сюда относятся забавные рассказы из жизни исторических лиц (так называемые исторические анекдоты), прибаутки, поговорки, афоризмы и т. д. Ср. два примера из книги [Раскин 2002: 145, 560] (заметим, что большинство приводимых в ней текстов вполне может быть рассказано как анекдоты)

(57) *У выдающегося ученого физика, лауреата Нобелевской премии Нильса Бора в кабинете висела подкова. — Неужели вы, великий физик, верите, что подкова приносит счастье? — Нет, не верю, — ответил Бор. — Но она приносит счастье независимо от того, верю я в это или не верю.*

[2] Особый тип письменной фиксации анекдота составляют записи «для памяти», когда записывающий стремится зафиксировать анекдот, чтобы впоследствии, обратившись к записи, вспомнить его и при случае рассказать другим. В записях «для памяти» чаще всего не фиксируется полный текст анекдота, а приводится лишь ключевая фраза или даже всего лишь несколько слов.

(58) *Как-то Крупская не стерпела какую-то очередную грубость Сталина и резко что-то ему ответила. — Патыше, таварищ Крупская! Если будэтэ сэбя плохо вэсти, мы назначим вдавой Лэнына Стасову.*

Ясно, что тексты такого рода было бы странно предварить «метатекстовым вводом» *Слыхал анекдот?*

3. «Протоанекдоты»

От письменной фиксации анекдотов следует отличать тексты, которые не являются записью устного рассказа, а изначально рассчитаны на читателя письменного текста. Такие тексты можно считать разновидностью «протоанекдотов», т. е. текстов, которые используются в рамках иных речевых жанров, а в речевом жанре рассказывания анекдота могут быть использованы лишь при условии необходимой языковой адаптации. К числу «протоанекдотов», предполагающих функционирование в устной форме, могут быть отнесены байки, устные новеллы, прибаутки, шуточные загадки. Нас в данной статье будут интересовать «протоанекдоты», функционирующие в письменной форме[3].

3.1. «Протоанекдоты» из «Сатирикона»

3.1.1. Текстовые «протоанекдоты»

Характерным примером таких «протоанекдотов» могут служить короткие смешные истории, опубликованные в журнале «Сатирикон», издававшемся в России в начале XX в.[4], т. е. в годы, когда речевой жанр анекдота еще только складывался. Возможно, такие истории появились вначале как подражание переводным историям из британских и американских журналов; эти истории публиковались в «Сатириконе» в разделах «Английский юмор» и «Английское остроумие», напр. (полужирным шрифтом выделены заголовки):

(59) **Жестоко срезал**. Посетитель (плешивый, волосы не растут выше ушей) *Постричься! Снять воротник?* Цирульник. — *О. нет, зачем беспокоиться! Можете остаться даже в котелке, если хотите.* (London Magazine)

(60) **Хиромант.** 1-ый босяк — *Я по линиям руки узнаю характер человека.* 2-ой босяк. — *А ну, какой у меня характер?* 1-ый босяк *(рассматривая руку). — Ты не любишь мыла.* (Chicago Tribune)

(61) **Учитель.** — *Что такое зебра?* — Ученик (живший летом на морском курорте)— *Это лошадь в купальном костюме.* (Punch)

«Протоанекдоты» такого рода в то время, насколько можно судить по публикациям в журнале, еще не назывались «анекдотами», но очевидно, что уже в первое десятилетие двадцатого века это был общепризнанный русский журнальный жанр. Эти «протоанекдоты» представляют собою авторские коротенькие рассказики: очень часто под такими историями стоит подпись. Чаще всего «протоанекдот» строится как обмен репликами (в предельном случае он состоит из единичной реплики)[5]. Авторы этих реплик (персонажи «протоанедота») могут быть названы в тексте (*он — она*, *лакей — посетитель*), но иногда никак не называются. Как правило, «протоанекдоты» снабжены заголовком, который задает общую ситуацию (*На уроке*; *На суде*; *В трактире*; *В кафешантане*; *На улице*) либо дает тот или иной комментарий (*Точное определение*; *Спасительная мера*; *Ответ по существу*; *Предусмотрительность*).

Частым приемом создания комического эффекта в «протоанекдотах» из «Сатирикона» является каламбур. Ср.:

(62) **На уроке.** Батюшка (подсказывая). — *Ну… на нравственность христианина имеет влияние среда…* Ученик (спохватываясь). — *И пятница как постные дни!*

---

[3] Некоторые типы «протоанекдотов», напр. исторические анекдоты, притчи и др., могут в равной мере успешно функционировать и в письменной, и в устной форме.

[4] Еженедельный журнал сатиры и юмора «Сатирикон», выходивший в Петербурге в 1908–1914 гг. под редакцией Аркадия Аверченко, был одним из самых популярных юмористических журналов начала двадцатого века. Авторы «Сатирикона» «позиционировали» себя как авторов нового для России типа юмористического журнала, созданного, как они сами отмечали в рекламе, «по образцу лучших немецких и английских сатирических изданий». Все примеры в статье взяты из подшивки журнала «Сатирикон» за 1909 г.

[5] Только небольшая часть «протоанекдотов», опубликованных в «Сатириконе», устроена несколько иначе, напр.: ***Объявление***. Требуются опытные мастера для починки расклеивающегося *кабинета*. Краснодеревцев просьба не приходить; ***Из дневника***. …Вот уже сорок дней прошло, как у меня с женой не было ни малейшей ссоры. Сегодня — сороковой день со дня ее смерти. Таев. Можно выделить еще несколько типов коротких юмористических текстов, публиковавшихся в журнале «Сатирикон»: пословицы и поговорки (У семи лидеров фракция без глазу; После реформ в четверг; Семь бед, один запрос. И. Гуревич; Погром гремит, а еврей не перекрестится. М.-Г.), афоризмы (***Стружки***. Некто говаривал: Человек обыкновенно бежит в двух случаях: если впереди него — хорошенькая женщина или — сзади злейший кредитор), загадки (***Тоже гаданье***. — Скажите, может ли, в конце концов, сделать что-нибудь премьер-министр в смысле реформ? — А ответ простой: стоит только прочесть подряд инициалы его имени, отчества и фамилии. [Внизу приводится ответ] П. А. С.).

(63) **На суде.** — *Почему вы, свидетель, называете подсудимого воришку банковским деятелем? — Я видел, как деятельно он таскал из кладовой — банки с вареньем.*

(64) — *Вы говорите, что вас спустили в одном доме с лестницы за то, что вы, будучи на блинах, громко восхищались икрой. — Да. — Странные хозяева. — Ничего не странные. Я восхищался икрой хозяйки…*

(65) **В ресторане.** Лакей. — *Какой крепости чаю прикажете?* Посетитель. — *Только не Петропавловской.* Голос с другого стола. — *И не Шлиссельбургской…*

«Протоанекдоты» из «Сатирикона» можно разделить на «бытовые» и «политические», хотя, конечно, это условное разделение, и бытовые темы часто переплетаются с политическими. Ср., напр.:

(66) **Тяжелый упрек**. *Русский беседует с французом: — Безобразие у вас, русских, делается… Воровство, казнокрадство, произвол, нищета!…*

Русский мрачно хмурит брови и молчит. А потом ядовито улыбается и обрушивается на француза: — А у вас… У вас зато… У французов… Неправильных глаголов в языке много!!!

Бытовые «протоанекдоты» свидетельствуют о том, что многие анекдотические стереотипы начали складываться еще в начале двадцатого века. Это стереотип властной жены, восприятие русских как пьяниц, стереотипное представление об американцах как о людях, ценящих свое время и деньги:

(67) — *Что с тобой?! Хромаешь, глаз подвязан!! — Пустяки. Это жена вчера спрашивала меня, где я был так долго.*

(68) **Под башмаком мужа**. — *Жена вас слушается? — О, беспрекословно! Вчера, например, она назвала меня идиотом. А, думаю — так! И приказываю ей: «а ну-ка, повтори еще раз!» И она, представьте, повторила еще раз. Волк*

(69) **Иностранец в Петербурге**. — *А у вас, повидимому, широко интересуются астрономией. — А что? — Где ни посмотришь, наводят на небо телескопы. — Какие телескопы? Это водку из горлышка пьют! Лик*

(70) **Американец**. — *Ты же говоришь, что любишь ее больше жизни? Когда же ты в нее влюбился? — В среду на прошлой неделе, в тридцать шесть минут двадцать семь секунд шестого вечера.*

В политических «протоанекдотах» из «Сатирикона» обыгрываются слова или поступки политических деятелей начала двадцатого века, представленные в Государственной Думе политические партии, политические события, происходящие в это время в России и в мире, издающиеся в Российской империи газеты, чуждого «Сатирикону» политического направления. Понимание многих политических «протоанекдотов» в настоящее время затруднено из-за отсутствия у современных носителей русского языка необходимых фоновых знаний. Так, истории, в которых упоминаются какие-то политические деятели, вовсе непонятны или не вполне понятны без знания того, кем были эти политики, а следующие два анекдота требуют некоторых знаний того, как расследовалось дело об убийстве Герценштейна:

(71) — *Так ты говоришь, что он клялся тебе в вечной любви — Да. По крайней мере, он говорил, что будет любить меня до тех пор, пока не кончится дело Герценштейна…*

(72) *Говорят, что, приехав в Ялту, Дубровин спросил у Думбадзе: — Ваше превосходительство! Депутата Герценштейна убили в Териоках, а мне пришлось ехать в Ялту. Теперь, если какого-нибудь депутата убьют в Ялте… то мне в Териоки придется ехать?*

(73) — *Вот убью я тебя, и ничего мне за это не будет — Ты что же… Слово такое знаешь? — Не слово, а просто сейчас же в Ялту уеду и в союзники поступлю. Суньсяка!*

Некоторые истории еще можно понять, додумывая бытующие в то время в русском обществе или внушаемые читателям авторами «Сатирикона» стереотипы (*союзники*, т. е. члены «Союза русского народа», — пьяницы, октябристы — мракобесы и непорядочные люди, Пуришкевич — зло для России и т. п.), но есть и истории, которые без знания исторических реалий или специальных комментариев вовсе непонятны, напр.:

(74) **География для кадетского корпуса**. — *Сколько верст от Выборга до Лондона? — Пустяк! Один неосторожный шаг.*

(75) **По нынешним временам**. — *Вы куда же сына определили? — В духовную семинарию. Что ж думаю, выучится, военным министром будет.*

Показательно (и это, конечно, не является случайностью) отсутствие в «Сатириконе» этнических «протоанекдотов»: этнические анекдоты, как и сексуальные, считались в то время неприличными и неудобными для печати. В нескольких номерах есть истории о евреях, но это, скорее, не этнические,

а политические «протоанекдоты» о черте оседлости, погромах или процентной норме (*проценте*); в них никак не отражены речевые и поведенческие особенности *еврея* как персонажа русского анекдота. Ср.:

(76) **Грамматика последнего дня**. *— Дети! Разберем простое предложение: еврей живет в Ялте. Где здесь подлежащее? — Конечно, еврей! — Почему — конечно!? — Потому что он — подлежащее высылке.*

(77) **На уроке истории**. *— Кто был князь Курбский при Иоанне Грозном? — Еврей. — Почему?!! — Потому что он не имел права жительства в России.*

(78) **В школе**. *— Исакович! Что такое — лютеранство — Ремесло. — Что за вздор? — …Потому что оно дает евреям правожительство.*

(79) **Имена**. *Один союзник спросил еврея: — Зачем вы, жиды, меняете еврейские имена на русские? — А вы? Ой! Черносотенцы еще хуже меняют русские имена на еврейские. Вас зовут Иван Пятерня, а после кишиневского погрома на ваших золотых часах стояло — Абрам Пинхусов, на платках — М. Янкелевич, а на наволоках — Сура Бескина!… Волк*

(80) **По привычке**. *В трамвае.* Еврейский юноша. *— Скажите, кондуктор… у вас здесь можно прямо садиться, или тоже три процента допускается?*

(81) **Точный ответ**. *На уроке арифметики. — Янкелевич, объясни мне, что такое — процент? — Процентом называется такое, от чего, если у Сидорова круглое три, он попадет в гимназию, а у Янкелевича если круглое пять, так ему говорят: пойди еще погуляй, Янкелевич… Волк*

Среди разнообразных отличий рассмотренных «протоанекдотов» от современных анекдотов более всего бросается в глаза наличие заголовка у большинства «протоанекдотов». Рассказывание современного анекдота не предполагает заголовка, и совершенно невозможен метатекстовый ввод: *Хочешь, я расскажу тебе анекдот под названием…* В то же время, устранив заголовок и проведши необходимую адаптацию, многие «протоанекдоты» можно рассказать как анекдоты в современном смысле. Укажем на соответствие некоторых «протоанекдотов» из «Сатирикона» и современных анекдотов:

(82) **Перед купанием**. *— Удивительно! Ты еще более грязен, чем я…— Что ж тут удивительного? Я же старше.*

Ср. современный анекдот: *Пошли Василий Иванович с Петькой купаться. Разделись. Петька говорит: «Смотри, Василий Иванович: у тебя ноги грязнее, чем у меня». — «Конечно, я же старше, Петя».*

(83) **В кофейной Филиппова**. *Двое за столиком. — Я выпью кофе по-варшавски. — А я по-одесски. — Это как же? — Такое же самое, только — заплатишь ты. Нуллюс*

Ср. анекдот: *Заходит еврей в публичный дом, и говорит: — Я хочу любовь по-еврейски. Бандерша говорит: — Я знаю любовь по-французски и прочие фокусы… Но любовь по-еврейски? В первый раз слышу! Потом одна из девочек говорит: — А я знаю. Пришли они в комнату, а девочка смущается: — Знаете, я вам соврала. Я не знаю любовь по-еврейски. Но дела у нас в борделе плохи, и если хотите, то вы можете иметь то же самое, но за полцены. Еврей обрадовался и говорит: — Так это же и есть любовь по-еврейски!*

(84) *— Ах! — нервно сказал интендант, — от этих ревизий у меня голова горит! — Не шапка ли? — спросил кто-то.*

Ср. современный анекдот: *Почему Лужков даже зимой ходит в кепке? — Да на нем шапка горит.*

В то же время «каламбурные» анекдоты плохо поддаются такой адаптации, несмотря на то, что каламбур остается важным средством создания комического эффекта и в современном анекдоте — однако способы построения каламбура стали несколько иными.

В целом, как мы уже отмечали [Шмелева, Шмелев 2005: 521–522], российская ситуация начала XX в. — это типичная ситуация свободного общества, в котором политический юмор является в основном принадлежностью газет и журналов, а в устном фольклоре по большей части бытуют анекдоты, которые не принято печатать в журналах: анекдоты неприличные или «неполиткорректные» анекдоты. Широкое распространение устных фольклорных анекдотов, появление большого числа политических фольклорных анекдотов в советское время было во многом обусловлено отсутствием свободной журналистики. Анекдоты стало принято рассказывать в интеллигентской среде — в среде тех, кто в другой ситуации являлись бы подписчиками таких журналов, как журнал «Сатирикон».

3.1.2. Рисунки с подписью

Распространенным типом «протоанекдота» в «Сатириконе» является также рисунок, сопровождаемый текстом. Текст чаще всего представляет собою речь лиц, изображенных на рисунке. Рисунки можно условно разделить на «бытовые» сценки и политические карикатуры. Приведем несколько примеров.

«Бытовые» сценки:

(85) (Рисунок: улыбающийся мужчина средних лет бреется, сидя перед зеркалом. Рядом стоит полуодетая женщина.) Муж.
— Когда я так вот утром побреюсь, то чувствую себя лет на 20 моложе. Жена (ласково). *— Послушай, мой милый, не лучше ли тебе бриться на ночь?!*

(86) (Рисунок: мужчина в спальне целует женщину.)
По опыту. *— Согласись, что ты рискуешь. Увез меня с маскарада в маске! Ну, а вдруг я — твоя жена? — Это немыслимо! Жена давно бы уже отобрала все мои деньги, укусила за руку, дала пару пощечин и в настоящий момент колотила бы по моей спине дождевым зонтиком!*

Рисунки на «политическую» тему:

(87) (Рисунок озаглавлен: ***Русские депутаты в Англии***. На рисунке изображены двое мужчин, уткнувшихся в словари, и британский полисмен.)
Отсталая страна. *— Черт его знает! Целый час по английскому словарю ищу, как по-ихнему «охранное отделение», — и даже намека на это слово нет.*

(88) (Рисунок: сидит толстый солидный господин, около него худой носатый человек.)
Учебная норма. *— Скажите, ваше благородие. Если мы на том свете попадем в ад — нас тоже 3 процента будут принимать?…*

(89) Два рисунка под общим заголовком: ***Как аукнется — так и откликнется***.

(90) (Рисунок 1: за столом с книгами и глобусом сидит человек в мундире. Около стола стоит в позе просителя бедно одетый человек с пейсами, который держит за руку маленького мальчика.) Директор гимназии. *— Конечно, можно. Но вы же знаете, что такое процент!*

(Рисунок 2: на диване сидит богато одетый человек с золотой цепью на животе. Перед ним в позе просителя стоит человек в мундире.) Банкир. *— Конечно, можно. Но вы же знаете, что такое процент!!*

Многие из таких «протоанекдотов» могут быть трансформированы в анекдоты и рассказаны в качестве таковых. Необходимая адаптация включает трансформацию изображения в текст «от повествователя». Так, пример (30) может быть рассказан при помощи следующих вербальных средств: *Разговаривают утром муж с женой…* или *Муж побрился утром и говорит жене:… А жена ему:…* Описание способов такой трансформации представляет собою нетривиальную и интересную семиотическую задачу.

3.1.3. Подача анекдотов в сборниках анекдотов

«Протоанекдоты» не прекратили свое существование и в советское время. Советские сатирические журналы (такие, как «Крокодил»), наряду с другими письменными юмористическими и сатирическими жанрами (такими, как юморески, фельетоны и т. п.) охотно печатали карикатуры на политические темы с подписями; часто печатались также карикатуры, высмеивающие «пережитки капитализма» (пьяниц, лодырей и т. п.) в советском обществе. Впрочем, «протоанекдоты» советской пропаганды с трудом поддавались адаптации к устному рассказыванию, и было не слишком вероятно, что кто-то стал бы их пересказывать в качестве анекдотов.

В то же время любопытно, что сборники анекдотов часто включают то, что с точки зрения оформления оказывается весьма близко к «протоанекдотам». В отношении немногочисленных сборников, выходивших в свободной печати за границей в период между мировыми войнами, это не удивительно: в то время анекдот как новый речевой жанр еще только формировался, так что в этих сборниках, наряду с анекдотами в «новом» смысле слова печатались исторические анекдоты, афоризмы, юмористические загадки, а анекдоты, как правило, снабжались заголовками. Но указанный обычай сохранился и после возрождения «анекдотопечатания»: анекдоты во многих из них снабжаются заголовками, а кроме того, как уже говорилось, ряд текстов в них представляют собою иные речевые жанры: так называемые исторические анекдоты, прибаутки, поговорки, афоризмы. Однако в современных сборниках анекдотов такая практика часто вызывает обманутые ожидания читателей, которые вправе рассчитывать, что в этих сборниках будет содержаться более или менее аккуратная фиксация анекдотов, действительно рассказываемых в обществе. Именно с этим, по-видимому, бывает связано отрицательная оценка таких сборников, присущая многим ценителям анекдотов.

В некоторых из таких сборников анекдоты сопровождаются иллюстрациями. Заметим, что функция таких иллюстраций отлична от функции рисунков с подписями, которые печатались и печатаются в юмористических журналах. Рисунок с подписью обычно устроен таким образом, что без рисунка подпись непонятна; в этом отношении рисунок заменяет необходимые комментарии «от повествователя». Рисунок и подпись составляют некое единое семиотическое целое. Что касается иллюстраций к анекдотам, они обычно используются исключительно для оживления: текст анекдота понятен и без них.

4. Анекдоты в Интернете

Интернет создал новую среду бытования анекдотов. В этой среде анекдоты представле-

ны в письменной форме, но таким образом, что их представление обнаруживает ряд черт устного рассказывания анекдотов. Более всего это касается функционирования анекдотов на форумах и чатах, когда записывание анекдота не отделено во времени от его восприятия. Анекдоты на форумах и чатах в норме сопровождаются «метатекстовыми вводами»; оформление текста анекдота бывает максимально приближено к его оформлению при устном рассказывании (напр., иногда используются «слова-паразиты»). Текст анекдота нередко существенно варьируется в конкретных актах реализации.

На размещенных в Интернете собраниях анекдотов представление текста анекдота в большей степени отличается от устного рассказа. В частности, на них не используются «метатекстовые вводы», и в этом отношении они подобны печатным сборникам анекдотов. С точки зрения приближенности к устному функционированию анекдотов уместно разграничить динамические и статические собрания (ср. [Крикманн 2004: 23]). Статические собрания, будучи созданными, не изменяются (или изменяются мало). К динамическим собраниям относятся многие сайты анекдотов, а также «отделы анекдотов» на порталах разного рода, побуждающие всех пользователей оценивать размещенные в собрании анекдоты, а также пополнять собрание. Варьирование текста анекдота в собраниях существенно меньше, нежели на форумах и чатах (которые А. Крикманн называет «контекстовыми» динамичными источниками — в отличие от динамических собраний, причисляемых им к «бесконтекстовым» динамическим источникам). Это и понятно: размещение анекдота в электронных собраниях часто осуществляется посредством копирования анекдота из уже имеющихся собраний. В тексте анекдотов из электронных собраний, как правило, отсутствуют показатели спонтанности (такие, как «слова-паразиты»).

В целом статические собрания с точки зрения подачи текста анекдота во многих отношениях близки к печатным собраниям анекдотов (некоторые статические собрания и представляют собою печатные собрания, переведенные в электронную форму), тогда как в динамических собраниях, равно как и на форумах и чатах, подача текста анекдота отличается некоторым своеобразием. Для них характерно стремление приблизить письменную форму текста к его устной форме, в частности передать интонацию в тех случаях, когда она релевантна. Для этого используются такие средства, как заглавные буквы, знаки препинания. С другой стороны, приближение к устной форме состоит в том, что в собственный текст анекдота включается только то, что произносит рассказчик анекдота; описание его интонации и жестов обычно подается не внутри текста анекдота, а при помощи ремарок (напр., в скобках), т. е. так, как это и следует делать при научной фиксации. Ср., напр., следующие четыре анекдота в том виде, в котором они были приведены на сайте www.anekdot.ru (который, понятно, относится к динамическим собраниям анекдотов): — *Как зовут глухого кролика? — КРОООООЛИИИИК!!!!!!!!!!!!*

(91) *Лежат два помидора в холодильнике. Один из них вдруг говорит: — Брррррррррр, как тут холодно!!! Второй ошалело на него посмотрел и как заорет: — Ааааааааааааааааааа!!!!!!!!!!!!! Говорящий помидор!!!!!!!!!!!!!!!*

(92) *В электричке, по проходу бежит маленькая девочка и поет — ля-ля ля-ля — (басом) МУЖИК НОГИ УБЕРИ! — ля-ля ля-ля.*

(93) *Гиви купил белый “Stainway”. Привез к себе в аул и дал концерт. Пришли все сельчане. Гиви вышел в белом фраке, сел за белый “Stainway”, открыл ноты и стал играть. Играл он одним указательным пальцем правой руки. Играл он долго… Час, два, три. Наконец он закончил. В комнате тишина. Через пять минут встает старец и говорит: «Гиви, нам очень понравилось. Но скажи… Мы смотрели по телевизору, показывали, как Рихтер играет… Так он играет всеми пальцами…» Гиви медленно встает и отвечает: «Рихтер, он еврей, он ищет (показывает как играет пианист). А я (показывает указательный палец, которым он играл) ЗНАЮ».*

Впрочем, в одном отношении текст анекдотов, приводимых в динамических собраниях, и отчасти анекдотов, рассказываемых на форумах и чатах, отличается от текста анекдотов в устном бытовании. Как мы уже неоднократно отмечали (в том числе в книге [Шмелева, Шмелев 2002]), стандартный порядок слов в начальном предложении анекдота — глагол, а затем подлежащее. В анекдотах из Интернета в значительном числе случаев порядок слов обратный: сначала подлежащее, а лишь затем глагол.

Отдельного замечания заслуживают способы, посредством которых в анекдотах Интернета цитируются сквернематерные выражения. При устном рассказывании анекдотов, содержащие такие выражения, они либо произносятся полностью, либо, если это возможно, стыдливо опускаются. В печатных сборниках анекдотов они иногда печатаются à toutes lettres (напр., в книге [Раскин 2002]), но чаще часть букв в них заменяется отточиями. Оба указанных способа используются и в Интернете, однако самым распространенным является третий способ, вообще характерный для русского Интернета, когда пропущенные буквы заменяются не отточиями, а особыми неконвенциональными символами.

5. Заключительные замечания

Итак, мы видим, что, обсуждая письменное бытование русских анекдотов, необходимо различать: фиксацию на письме анекдота, который был рассказан или должен быть рассказан устно; тексты, которые не могут быть рассказаны как анекдоты без более или менее значительной трансформации («протоанекдоты»; юмористические картинки и карикатуры с подписями); анекдоты, бытующие и в каком-то смысле «рассказываемые» в Интернете. Указанные типы письменных текстов, соотносимых с анекдотами, различаются тем, как в них взаимодействуют письменная форма текста и его устное воспроизведение при его рассказывании в рамках речевого жанра анекдота.

Фиксация на письме текста анекдота представлена в основном в печатных сборниках анекдотов, а также в электронных собраниях Интернета (особенно статических). «Протоанекдоты» не могут рассматриваться как анекдоты в собственном смысле слова; для них письменная форма бытования является основной, а естественной средой бытования являются разного рода юмористические журналы. Они могут включаться в печатные сборники анекдотов, а также быть представлены в электронном виде в Интернете, но такая форма бытования для них вторична. Можно было бы ожидать, что расширение доступного на письме арсенала изобразительных средств, явившееся следствием развития электронной письменной коммуникации разных видов, приведет к тому, что протоанекдоты обретут свою естественную среду бытования в Интернете; однако пока этого не произошло. Интернет в качестве среды бытования интересен тем, что привел к особой форме письменного «рассказывания»; текст, с которым мы имеем дело в таком случае приобретает некоторые черты устного «рассказывания», включая показатели спонтанности.

Литература

1. *Вознесенский А. В.* О современном анекдотопечатании // Новое литературное обозрение, 1996. С. 393–399.
2. *Крейдлин Г. Е., Шмелева Е. Я.* Вербальные и невербальные элементы анекдота // Логический анализ языка. Языковые механизмы комизма М.: 2007. С. 509–519.
3. *Крикманн А.* (составитель и редактор) Интернет-анекдоты о Сталине. // Тарту, 2004.
4. *Раскин И.* (составитель) Анекдоты от Иосифа Раскина. // М.: ПБОЮЛ Малиновская, 2002.
5. *Шмелева, Шмелев 1998* — Шмелева Е., Шмелев А. Рассказывание анекдота как жанр современной русской устной речи// Труды международного семинара «Диалог'98» по компьютерной лингвистике и ее приложениям. Казань, 1998. С. 262–271.
6. *Шмелева Е. Я., Шмелев А. Д.* Русский анекдот. Текст и речевой жанр. // М.: Языки славянской культуры, 2002.
7. *Шмелева Е. Я., Шмелев А. Д.* Русский анекдот в двадцать первом веке: трансформации речевого жанра // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции Диалог'2005. М., 2005.

Что может помочь компьютеру понять, кто стоял на балконе

What can help the computer to learn who was on the balcony

Юдина М. В. (maria_yu@abbyy.com)
ABBYY

В статье перечислены некоторые факторы, влияющие на разрешение синтаксической неоднозначности, с точки зрения возможности их использования в автоматическом анализе текста, а также показаны результаты опыта семантического подхода к разрешению синтаксической неоднозначности.

Одну из наибольших проблем для автоматической обработки текста составляет синтаксическая омонимия, или синтаксическая неоднозначность, т. е. возможность построить несколько синтаксических структур на основе одной и той же последовательности языковых знаков. В подавляющем большинстве случаев омонимию может разрешить только человек. В ряде случаев омонимия принципиально неразрешима без дополнительной информации (например, в предложении (1): *Маша читала и писала письма*, см. [Юдина, Янович, Фёдорова 2007]). Тем не менее, в реальной жизни мы очень редко замечаем синтаксическую омонимию, в силу способности нашего синтаксического парсера мгновенно анализировать не только синтаксическую структуру, но и ситуацию, контекст, делать логические выводы о смысле предложения. Научить этому машинный анализатор представляется практически невозможным.

Изучение синтаксической омонимии сводится, как правило, к исчислению всех потенциально возможных конструкций (см. например, [Иорданская 1967]). В системах автоматического анализа текста такая омонимия (если набор эвристик конкретного парсера в принципе может распознать некоторые типы омонимии), разрешается в пользу одного из вариантов случайным образом или с помощью статистики.

Рассмотрим один из наиболее изученных и популярных типов синтаксической неоднозначности — «раннее-позднее закрытие», или омонимию относительного придаточного предложения (см., к примеру, обзор в [Фёдорова, Янович 2004]). Напомним, что данная омонимия заключается в наличии двух интерпретаций для предложения (2) *Преступник застрелил служанку актрисы, которая стояла на балконе*: первое прочтение, называемое «ранним закрытием» (далее РЗ), соответствует пониманию «служанка стояла на балконе», а второе прочтение — «позднее закрытие» (далее ПЗ) — пониманию «актриса стояла на балконе».

Феномен раннего-позднего закрытия широко исследован на разных языках, однако на данный момент все полученные данные носят лишь теоретический характер, не находя применения в системах автоматической обработки текста и машинного перевода. Остановимся на некоторых факторах, влияние которых на выбор той или иной интерпретации в рамках разрешения омонимии раннего-позднего закрытия представляется доказанной. Данные факторы, как кажется, в перспективе могли бы быть использованы и в целях автоматического разрешения неоднозначности.

1. Длина придаточного предложения

Влияние длины придаточного предложения на выбор РЗ или ПЗ было проведено в работе [Fodor 1998]. Как оказалось, английский язык, в котором предыдущие эксперименты выявили предпочтение ПЗ, следует данному принципу далеко не всегда. Например, если придаточное предложения было длинным, ПЗ встречалось редко или не наблюдалось вовсе; иная картина наблюдалась в случае придаточных, состоящих из одного просодического слова: ПЗ было преобладающим. Данный факт был

проверен также на материале других языков: арабского, хорватского, французского, немецкого и испанского. Результаты оказались схожими: короткие придаточные гораздо чаще модифицировали второе существительное именной группы (далее ИГ). Для объяснения данной закономерности Дж. Фодор выдвинула Закон антигравитации. Данный закон гласит, что если присоединяемая составляющая имеет просодически «легкий» статус, то она, скорее всего, присоединяется к зависимому существительному ИГ; в то время, как присоединение более «тяжелых» составляющих зависит от двух факторов: от соотношения «просодической тяжести» придаточного предложения и ИГ, которую оно призвано модифицировать, и от просодических особенностей данного языка. Результаты были проверены и на русском материале в работе [Фёдорова, Янович 2004], подтвердившей результаты Фодор.

Фактор длины придаточного наиболее легко реализуем в системах автоматического синтаксического анализа.

2. Лингвистическая настройка.

Необходимо также упомянуть о Гипотезе лингвистической настройки (*Linguistic Tuning*), рассматриваемой в работе [Mitchell et al. 1995] (см. также [Драгой 2006]). Данная гипотеза предполагает, что одним из факторов, влияющих на разрешение синтаксической неоднозначности, является предыдущий лингвистический опыт человека в разрешении подобной неоднозначности. В частности, это означает, что выбор интерпретации предложения основан на той стратегии, которая является наиболее частотной в конкретном языке. Эта гипотеза была подтверждена на французском и английском материале.

Гипотеза лингвистической настройки может служить опорой для статистического подхода к разрешению синтаксической неоднозначности.

3. Одушевленность существительных, входящих в ИГ

Исследование [Brysbaert, Mitchell 1996], проведенное для проверки гипотезы лингвистической настройки на материале нидерландского языка, показало прямо противоположные результаты. В работе [Desmet et al. 2002] приводится объяснение этому факту. Было доказано, что при наличии в ИГ одушевленного и неодушевленного существительных одушевленное существительное выбирается чаще, чем неодушевленное в любой позиции, при наличии в ИГ двух одушевленных существительных преобладает РЗ, а при наличии двух неодушевленных — ПЗ; таким образом, исследование [Brysbaert, Mitchell 1996], в экспериментальном материале которого предложения с одушевленными именами составляли меньшинство, не может служить опровержением Гипотезы лингвистической настройки.

Фактор одушевленности/неодушевленности также достаточно просто учесть при автоматическом разрешении неоднозначности.

4. Контекст

Первым экспериментом, исследовавшим влияние контекста на присоединение придаточных предложений и проведенным по методике заканчивания предложений, стала работа Т. Десмета и коллег ([Desmet, De Baecke et al. 2002]). Ими был проведен эксперимент на базе нидерландского языка. Экспериментальный материал состоял из 30 предложений, к каждому было придумано по три контекста. Контекст, склоняющий к первому (главному) существительному ИГ, вводил двух возможных референтов для главного существительного и одного возможного референта для зависимого существительного ИГ; контекст, склоняющий ко второму, зависимому, имени, наоборот, вводил двух возможных референтов для зависимого имени и одного возможного референта для главного имени; и, наконец, в нейтральном контексте референты либо не вводились вовсе, либо вводился лишь один возможный референт для обоих существительных. Гипотеза авторов состояла в том, что, поскольку нидерландский язык относится к языкам с тенденцией к РЗ, нейтральный контекст будет способствовать присоединению относительного придаточного к вершине ИГ. При контексте, склоняющем к вершине именной группы, РЗ будет преобладать, а при контексте, склоняющем к зависимому существительному ИГ, будут возможны случаи ПЗ. Результаты эксперимента полностью подтвердили гипотезу.

Серия подобных экспериментов на русском материале была проведена в работе [Юдина 2006]. Всего в исследовании приняли участие 165 человек, было обработано 2970 экспериментальных предложений. Общие результаты экспериментов таковы: 62 % РЗ после нейтрального контекста, 91 % РЗ после контекста, склоняющего к первому имени ИГ, 60 % РЗ после контекста, склоняющего ко второму имени ИГ.

Фактор контекста на данном этапе его изученности наиболее сложен для формализации с целью использования его для автоматического разрешения неоднозначности, так как помимо референциального аспекта введения участников неминуемо содержит также и семантическую информацию о ситуации в целом, что, несомненно, также влияет на тип закрытия.

5. Прайминг

Явление синтаксического прайминга заключается в следующем: при ответной реакции на какой-либо стимул говорящий склонен использовать те синтаксические конструкции, которые он в недавнем прошлом каким-либо образом обработал (услышал, прочитал, сказал). Одним из типичных проявлений синтаксического прайминга является синтаксическая координация участников диалога. Высказывание, осуществляющее преднастройку, называется «праймом», а высказывание, на порождение или понимание которого, как предполагается, окажет влияние прайм, называют «целью». Первый подобный эксперимент был проведен в [Scheepers 2003] на немецком материале. Экспериментальный блок состоял из четырех предложений: прайм с ранним закрытием (далее РЗ-прайм), прайм с поздним закрытием (далее ПЗ-прайм), базовый прайм (блокировал присоединение относительного придаточного) и целевое предложение (далее цель). Суть эксперимента заключалась в следующем: РЗ- и ПЗ-праймы могли быть продолжены испытуемыми только одним способом, базовый прайм не предполагал конструкции с относительным придаточным, а цель была составлена так, что она могла быть закончена испытуемыми двояко. Предполагалось, что вынужденное использование испытуемым РЗ или ПЗ в прайме вызовет использование соответствующей структуры в синтаксически неоднозначном предложении — цели. Базовые праймы включались в экспериментальные наборы для того, чтобы проверить, каково будет предпочтение закрытия в цели в отсутствие прайминга. Результаты эксперимента показали значительный прайминг-эффект.

Мы провели серию аналогичных экспериментов на русском материале ([Юдина, Фёдорова 2009]). Всего было проведено 3 эксперимента, в которых приняли участие 131 человек, было обработано 2430 экспериментальных предложений. Основной эксперимент серии показал значительный эффект синтаксического прайминга: 57 % РЗ после РЗ-прайма и 46 % РЗ после ПЗ-прайма. Количество случаев РЗ после базового прайма составило 60 %, что подтвердило результат предыдущих экспериментов.

Влияние эффекта прайминга на разрешение неоднозначности имеет больше теоретическую значимость, нежели практическую; однако, наличие не только глобальной, но и локальной настройки на разрешение неоднозначности определенным способом также может помочь более тонкому и правильному статистическому подходу к разрешению неоднозначности.

Таким образом, после проведения экспериментов, исследовавших влияние прайминга и контекста, в нашем распоряжении оказалась огромная база экспериментальных предложений с разрешенной неоднозначностью раннего-позднего закрытия. При обобщении результатов всех проведенных экспериментов можно выявить среднее для русского языка распределение типов закрытия в отсутствие какого-либо склонения к одному из типов закрытия, будь то контекст или прайминг. Мы считаем, что это распределение равно 60/40 %, что вполне согласуется с другими экспериментами с ранним-поздним закрытием на русском материале (см., например, [Sekerina 2003]). Принятие этих цифр за некий базовый коэффициент дает возможность более подробно исследовать предложения, явно выбивающиеся из этих показателей, или же наоборот, выявить наиболее «стандартные» с точки зрения закрытия предложения.

6. Тип глагольной вершины (предикат главного предложения)

Все проведенные нами эксперименты были основаны на приблизительно одном и том же экспериментальном материале. Некоторые предложения, казавшиеся нам неудачными, мы заменяли другими, менялся экспериментальный дизайн, но большая часть предложений оставалась неизменной. В результате мы оказались обладателями объемной базы экспериментальных предложений, при этом все они имели однотипную структуру: *Subject Verb* NN_{gen} *[Relative Clause]*.

Оказалось, что некоторые предложения показывают схожие результаты: например, большее по сравнению с нормальным количество случаев РЗ или ПЗ. В части случаев виной тому, несомненно, был общий контекст предложения, обрисовывавший ситуацию, заведомо так или иначе связанную с одним из существительных ИГ в качестве участника. Так, большинство контекстов, оказавшихся склоняющими к ПЗ, содержат концепт того же семантического поля, что и второе существительное ИГ (например, (3) *Редакция поручила Косте написать статью о шоу-бизнесе, но он совершенно не знал, с чего начать. Коля попросил помощи у менеджера певицы,…*: «певица» относится к семантическому полю «шоу-бизнес»), или концепт, тесно связанный со вторым существительным тематически (например, (4) *Вся страна со страхом наблюдала за развитием событий в захваченном здании. Власти обещали освободить заложницу террориста,…*).

Однако если не брать в расчет контекст, а оперировать лишь непосредственно предложениями, оказывается, что одинаковым поведением обладают предложения с одинаковым типом глаголов-вершин. Например, для закрытия экспериментального предложения (5) *В парке друзья встретили ассистентку профессора,…* испытуемые предпочитали присоединение придаточного предложения к первому имени, что было подтверждено результатами двух экс-

периментов. Мы предположили, что ИГ, состоящая из одушевленных существительных разного рода, неравноправна с точки зрения закрытия, а именно, что существительные женского рода обладают свойством «перетягивать» закрытие. Для проверки этого факта мы провели дополнительный эксперимент, в котором поменяли род у существительных ИГ (например, «мама однокурсника» вместо «папа однокурсницы»), однако, этот эксперимент показал слабое влияние женского рода (в основном свойством притягивания закрытия обладали маркированные слова вроде «директриса»). Большинство предложений, несмотря на изменение рода существительных, показали результаты, полностью идентичные основному эксперименту.

Поэтому мы предположили, что существует и еще один фактор, влияющий на разрешение синтаксической неоднозначности, а именно, тип глагольной вершины. Из всех отобранных нами экспериментальных предложений мы составили список глагольных вершин, встречающихся более чем в одном экспериментальном предложении каждого эксперимента в отдельности или по всем экспериментам в целом. Всего были обработаны материалы четырех проведенных нами в разные годы экспериментов. Поскольку наши эксперименты имели разный дизайн и разные задачи (в некоторых экспериментах предложение функционировали со своими контекстами, в некоторых — без), можно считать, что одинаковые результаты, показанные предложением, не связаны с контекстом или другим влиянием.

Таким образом были выделены следующие глаголы:

Заметить
Увидеть
Смотреть
Слушать
Встретить
Навестить
Столкнуться
Разругаться
Поссориться
Договориться
Познакомиться
Узнать

Мы составили таблицу, иллюстрирующую количество РЗ и ПЗ в предложениях, содержащих данный глагол. Критерием оценки стало количество РЗ при нейтральном контексте (или после базового прайма) и при позднем контексте (или после прайма, склоняющего к ПЗ). Мы считали, что предикат способствует РЗ, если число РЗ в вышеуказанных условиях составляло 100–80 %, и к ПЗ — если меньше 50 %.

Далее, каждый предикат был отнесен к определенному типу согласно тезаурусу Роже ([Roget's Thesaurus 2000]). Мы выделили несколько групп однотипных предикатов; некоторые предикаты, такие, как «нагрубить» или «разговориться», пришлось не учитывать в исследовании, так как в рамках выборки они оказались единственными предикатами своего типа.

Результаты можно увидеть в таблице 1.

Наибольшую сложность вызвал предикат «увидеть». Дело в том, что во всех экспериментальных предложениях, кроме одного, этот глагол показывал большой процент ПЗ. Оказалось, практически во всех экспериментальных предложениях глагол «увидеть» означал скорее «увидеть и узнать», например (6) *Выходя на улицу, Катя вдруг увидела ученицу доктора, …* (увидела и узнала, что это именно ученица доктора). Лишь в предложении (7) *В конце концов Алексей увидел посетительницу директора, …* «увидеть» функционирует в своем прямом значении глагола зрения. Обозначим его «увидеть1», а другое

Таблица 1

предикат	тип закрытия	класс предиката по Роже
заметить	Р	matter → organic matter → vision
смотреть	Р	matter → organic matter → vision
слушать	Р	matter → organic matter → hearing
увидеть1	Р	matter → organic matter → vision
встретить	Р	space → motion → arrival
навестить	Р	space → motion → arrival
столкнуться	Р	space → motion → arrival
разругаться	П	volition → individual volition → antagonism → dissention
поссориться	Р\П	volition → individual volition → antagonism → dissention
договориться	Р	volition → individual volition → antagonism → concord
познакомиться	П\Р	intellect → formation of ideas → results of reasoning → knowledge
узнать	П	intellect → formation of ideas → results of reasoning → knowledge
увидеть2	*П*	*intellect → formation of ideas → results of reasoning → knowledge*

значение будем маркировать как «увидеть2» (выделено в таблице курсивом). Однако в словарях не фиксируется подобное различие смыслов; результаты, показанные предложением «Алексей увидел…» могут быть связаны с особенностями ИГ «посетительница директора»: например, причина может быть в разных референтных статусах ИГ «посетительница директора» и «ученица доктора». Данный вопрос подлежит дальнейшему изучению.

Как видно из таблицы, глаголы одного класса показывают удивительное единство в том, как разрешается синтаксическая неоднозначность в предложении, где вершиной является данный предикат. А именно: предложения с глаголами чувственного восприятия и глаголы движения имеют большой процент РЗ, глаголы, связанные с мышлением и интеллектом — ПЗ. Класс глаголов отношений неоднороден: глагол «договориться» показывает практически 100 % РЗ. Как нам кажется, это может быть связано со структурой ИГ: экспериментальное предложение (8) *Я решил договориться с женой строителя, …* почти все испытуемые заканчивали описанием того, что может сделать с (нерадивым) строителем жена (запретит пить, будет бить каждый вечер и т. п.)

Также не совсем ясна ситуация с глаголом «познакомиться»: обычно этот глагол ведет себя по варианту ПЗ, но одно предложение в двух экспериментах показало почти 100 % РЗ: (9) *Вчера Татьяна наконец познакомилась с секретаршей начальника, ….* Видимо, это также связано с особенностью самого этого предложения, точнее, с неудачной структурой ИГ с точки зрения семантики. Вероятно, «секретарша начальника» — это настолько устойчивый концепт, что опознаётся скорее как одно целое, чем как два независимых объекта.

Не вошедшие в нашу выборку по причине своей «единичности» в рамках экспериментальной выборки или в рамках классификации по Роже глаголы (например, *нагрубить, попрощаться*) также, тем не менее, показывают довольно однородные результаты. Нам кажется, что имеет смысл продолжить исследования в этом направлении.

Конечно, данные результаты нельзя считать окончательными и доказанными. Но они, на наш взгляд, иллюстрируют возможность семантического подхода к синтаксической неоднозначности. Класс предиката хорошо формализуем, поэтому не составит труда встроить семантическую классификацию предикатов в синтаксический анализатор.

Литература

1. *Юдина М. В., Федорова О. В., Янович И. С.* Синтаксическая неоднозначность в эксперименте и в жизни // Диалог. М.: 2007
2. *Иорданская Л. Н.* Синтаксическая омонимия в русском языке (с точки зрения автоматического анализа и синтеза) // НТИ. 1967. № 5.
3. *Фёдорова О. В., Янович И. С.* Об одном типе синтаксической многозначности, или Кто стоял на балконе. // МГУ, 2004
4. *Fodor J. D.* Learning to parse? // Journal of Psycholinguistic Research, 27, 2, 1998.
5. *Mitchell D. C., Cuetos F., Corley M. M. B. & Brysbaert M.* Exposure-based models of human parsing: Evidence for the use of coarse-grained (nonlexical) statistical records. // Journal of Psycholinguistic Research, 24, 6, 1995.
6. *Драгой О. В.* Разрешение синтаксической неоднозначности: правила и вероятности. // Вопросы языкознания, № 6, 2006
7. *Brysbaert M., Mitchell D. C.* Modifier attachment in sentence parsing: Evidence from Dutch. // Quarterly Journal of Experimental Psychology, 49A, 3, 1996.
8. *Desmet T., Brysbaert M. & De Baecke C.* The correspondence between sentence production and corpus frequencies in modifier attachment. // Quarterly Journal of Experimental Psychology, 55A (3), 2002.
9. *Desmet T., De Baecke C., Brysbaert M.* The influence of referential discourse context on modifier attachment in Dutch. // Memory & Cognition 2002, 30 (1), 2002
10. *Юдина М. В.* Понимание и порождение высказываний с синтаксической неоднозначностью (на примере относительных придаточных в русском языке) // Диалог. М.: 2006
11. *Scheepers C.* Syntactic priming of relative clause attachments: persistence of structural configuration in sentence production. // Dundee: University of Dundee, 2003
12. *Юдина М. В., Фёдорова О. В.* Разрешение синтаксической неоднозначности: эффекты прайминга и самопрайминга. // Диалог. М.:2009
13. *Sekerina I.* The Late Closure Principle in Processing of Ambiguous Russian Sentences. // The Proceedings of the Second European Conference on Formal Description of Slavic Languages. Universität Potsdam, Germany. 2003
14. *Roget's* Thesaurus Of English Words And Phrases // Penguin Books, London, 2000.

Просодия предложений со «снятой» иллокутивной силой[1]

Prosody of sentences with no illocutionary Force

Янко Т. Е. (tanya_yanko@list.ru)

Институт языкознания им. В. В. Виноградова РАН

В произведениях устной речи, не отражающих стандартных иллокутивных противопоставлений, т.е. не различающих частотных контуров сообщений, императивов и вопросов, просодия служит для сегментации на слова, предложения, тексты. Анализируется православное литургическое чтение, мусульманская молитва, чтение Бродским его собственных стихов.

В устной речи основная функция просодии — выражение иллокутивных сил. Так, в русском языке только просодия отличает сообщение (повествовательное предложение) от *да-нет*-вопроса, а высказываниям, где имеются сегментные средства выражения иллокутивной силы, например наклонение в императиве, тоже свойственны специальные суперсегментные показатели — определенные просодические контуры, которые выражают иллокутивные значения совместно с сегментными маркерами. Между тем существуют особые психологические состояния человека и определенные устные жанры, в которых иллокутивная функция просодически остается не выраженной. Так, не выражается просодически иллокутивная сила в пении, различных видах речитатива, чтении стихов и священных текстов в некоторых литургических традициях. Когда человек поет, мелодия нивелирует языковые движения тона. Однако мелодекламация, литургическое чтение, чтение поэтами собственных стихов пением в чистом виде не являются, и некоторые коммуникативные значения здесь все-таки присутствуют. Возникает вопрос, какие это значения и как конкретно они выражаются.

Указания на связь между музыкальной организацией речитатива, в частности, церковных песнопений, и языкового синтаксиса традиционно служили предметом анализа в работах искусствоведов и музыковедов (см., например, Владышевская 1983, Кутузов 1999, Пантелеева 2009 и цитированную в этих работах литературу). Так, в работе Т. Ф. Владышевской (1983), посвященной анализу музыкально размеченных древних текстов, высказывается гипотеза о возможном соответствии между знаками так называемой экфонетической нотации (указаниями, какие слова и как выделять при распевном чтении ритуальных текстов) и знаками препинания. Однако конкретных соответствий между экфонетической нотацией и языковыми единицами, такими, как конец предложения, конец текста и иллокутивными значениями не устанавливаются. С другой стороны, анализ просодических (прежде всего, частотных) особенностей современного чтения православных молитв и произнесения проповедей с разбором по жанрам и традициям дан в книге О. А. Прохватиловой (1999), где показано, в чем состоят средства выражения иллокутивных значений в некоторых православных традициях и в каком отношении к просодии спонтанной речи эти средства находятся.

Цель данной работы состоит в анализе функций просодии в произведениях речи, в которых не выражена иллокутивная сила, и в установлении сходств и различий между речью со снятой иллокутивностью и обычной речью. Кроме литургического чтения, будет рассмотрено произведение иного жанра, который тоже характеризуется отсутствием выражения иллокутивных сил: это чтение Иосифом Бродским его собственного стихотворения.

Ниже представлены результаты анализа нескольких конкретных чтений, которые в той или иной степени представляют соответствующую тра-

[1] Работа над темой финансируется Российским Государственным гуманитарным фондом, проект N08-04-00165а.

дицию или стилистику. Окончательных выводов о просодической структуре традиций чтения при таком подходе мы не делаем, но, как кажется, даже анализ отдельных произведений звучащей речи позволяет пролить свет на феномен «снятой иллокутивности».

При анализе плана выражения просодии основным параметром служат частотные показатели, которые отражают релевантные подъемы и падения частоты основного тона речи. Второй релевантный показатель — выбор словоформ-носителей акцентных пиков: изменения частоты значимы не сами по себе, а в связи с тем, на какой словоформе в предложении они фиксируются. Насколько нам известно, в работах, посвященных просодии литургической речи (а часто и речи вообще) параметр выбора словоформ-носителей акцентных пиков не учитывается. В данной работе мы надеемся восполнить этот пробел. Третий релевантный показатель это длительность ударного слога словоформы-носителя акцентного пика и заударных слогов.

Мы выделяем три уровня анализа материала. Первый уровень — анализ физической стороны речи. Релевантные подъемы и падения частоты основного тона, если они есть, нужно распознать на слух (и/или выделить на графике частот) и отличить их от нерелевантных. Второй уровень состоит в собственно лингвистической интерпретации наблюдаемых явлений: какие иллокутивные и другие коммуникативные значения можно увидеть за изменениями тона, т.е. какой план содержания можно приписать тонально-темпоральным параметрам чтения ? Третий уровень интерпретации состоит в анализе функций наблюдаемых явлений в культурной — литургической или поэтической — традиции. Почему священники и поэты именно так интонируют свою речь? Какая иллокутивная сила высшего порядка замещает отсутствующие стандартные иллокуции? Большинство работ, посвященных проблемам анализа просодии церковной речи, например, фундаментальный труд О. А. Прохватиловой (1999), ставят соответствующую задачу и выдвигают интересные гипотезы. В нашей работе этот вопрос останется за рамками исследования. Мы сосредоточимся только на проблемах формального разбора звучащих произведений речи.

1. Православное литургическое чтение

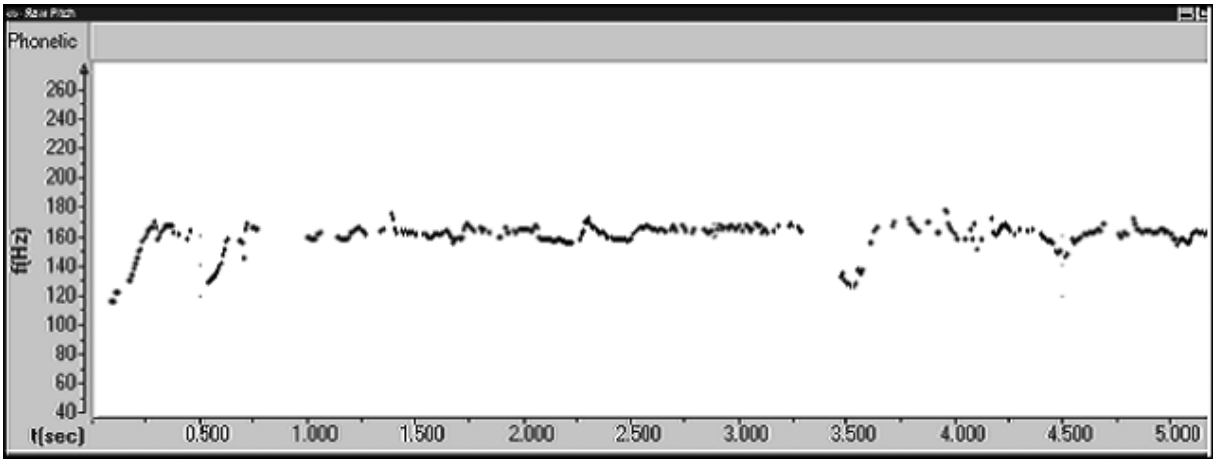
Начнем анализ с сопоставления записей литургического чтения православным священником молитвы «Отче наш» и речи того же человека в жанре наставления духовного отца пастве, которое делается в относительно свободной форме. Текст молитвы «Отче наш» приведем полностью.

Отче нашъ, Иже еси на небесѣхъ, (I)
да свѧтится имѧ Твое, (II)
да пріидетъ царствїе Твое: (III)
да будетъ волѧ Твоѧ, яко на небеси и на земли, (IV)
хлѣбъ нашъ насущный даждь намъ днесь, (V)
и остави намъ долги нашѧ, (VI)
ѧкоже и мы ѡставлѧемъ должникомъ нашымъ, (VII)
и не введи насъ во искушенїе, (VIII)
но избави насъ от лукаваго. (IX)

Графики (тонограммы), приведенные ниже, фиксируют изменение частоты основного тона речи. Ось абсцисс отражает время в секундах, а ось ординат — частоту в герцах. На Тонограмме 1 представлен фрагмент молитвы (строки (IV) и (V)), на Тонограмме 2 — фрагмент беседы священника с верующими.

(94) *… да будетъ волѧ Твоѧ, яко на небеси и на земли, хлѣбъ нашъ насущныи даждь намъ днесь*

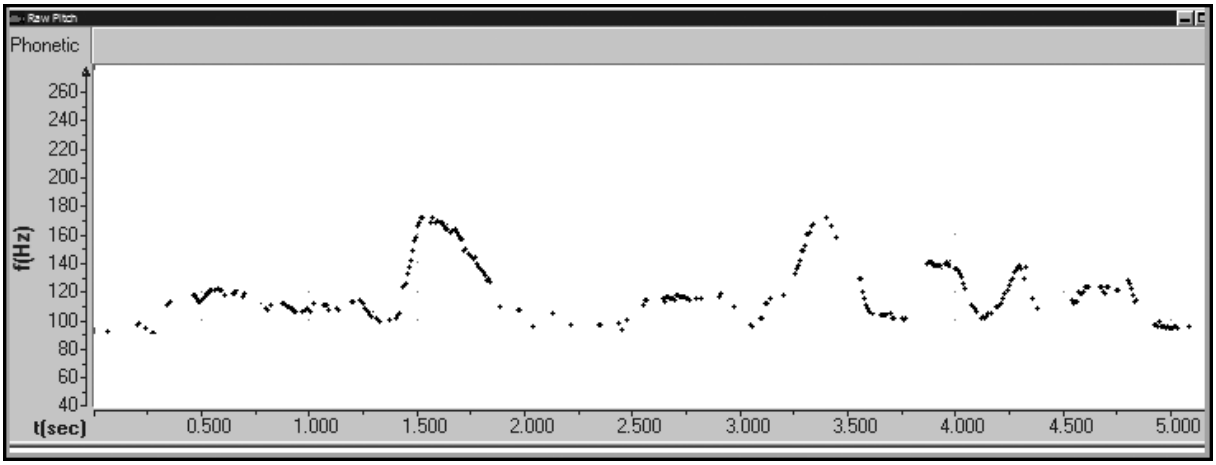
Тонограмма 1



да будетъ волѧ Твоѧ, яко на небеси и на земли, хлѣбъ нашъ насущныи даждь намъ днесь
(IV) (V)

(95) *…для того чтобы молитвы наши в наших устах звучали не лицемерно…*

Тонограмма 2



… для того чтобы молитвы наши в наших устах звучали не лицемерно…

Сравним два графика. Говорит один и тот же человек. Как эксплицировать общие просодические различия, которые здесь наблюдаются?

Тонограммы показывают, что молитва в целом звучит на более высоком среднем уровне. Минимальная частота молитвы — 120 герц, минимальная частота обычной речи данного говорящего — около 90 герц.

1. Диапазоны частот подъемов в обычной речи существенно больше, чем в литургической.
2. В примере (2) после словоформы *наши* в именной группе *молитвы наши* имеется существенная пауза. Эта пауза отделяет первую тему предложения *Для того чтобы молитвы наши* от второй темы *в наших устах*. Между темой *в наших устах* и ремой *звучали не лицемерно* (отрицательная частица *не* здесь просодически выделена говорящим) тоже имеется небольшая задержка артикуляции. При литургическом чтении паузы между началом и концом строки фактически отсутствуют.
3. Темы в предложении (2) *Для того чтобы молитвы наши* и *в наших устах* отмечены типичным акцентом темы с подъемом на ударных слогах словоформ *молитвы* и *устах* и падением на заударном слоге словоформы *молитвы*. Это акцент типа интонационной конструкции ИК-3 (по Е. А. Брызгуновой, об интонационных конструкциях см. Русская грамматика 1982: 103–118). На ударном слоге акцентноносителя ремы *не лицемерно* можно наблюдать легкое падение частоты тона, которое продолжается и на заударных (ИК-1, по Брызгуновой), что полностью соответствует маркеру ремы в русском языке. Сами акцентоносители — *молитвы*, *устах* и *не лицемерно* — выбраны в полном соответствии с принципами выбора акцентоносителя в коммуникативном компоненте предложения, таком как тема, рема, компонент вопроса или императива. Отдельно на принципах выбора акцентноносителя мы здесь не останавливаемся, они изложены в Янко 2001: 68–84; 2008: 38–60[2].
4. В молитве падений нет. Есть только подъемы тона на ударных слогах с высокими ровными заударными, когда на заударных слогах частота держится на том уровне, который был достигнут в результате первоначального подъема на ударном слоге. Ровный тон в данном случае служит характеристикой распевного чтения.

[2] По поводу словоформы-акцентоносителя поясним, что ее выбор определяется актантной структурой предиката, активацией, исключающей имя референта, названного в предтексте, из списка претендентов на роль акцентоносителя, внутренней структурой именных, глагольных и нексусных групп, которые могут представлять собой атрибутивные и сочиненные группы и в которых действуют внутренние правила выбора акцентоносителя, идиоматичностью заполнения валентностей и другими факторами. В результате действия этих факторов в языке вырабатывается определенная иерархия синтаксических компонентов предложения в соответствии с их готовностью служить акцентоносителем коммуникативного компонента предложения. Так, второе дополнение имеет приоритет перед первым дополнением и подлежащим; подлежащее имеет приоритет перед финитной формой глагола; несогласованное определение имеет приоритет перед определяемым именем; определяемое имеет приоритет перед согласованным определением. Например, предложение *Ветер* **ветку** *клонит* с акцентоносителем словоформой *ветку* иллюстрирует приоритет дополнения над подлежащим.

В результате можно предположить, что в каноническом литургическом чтении членение на темы и ремы отсутствует, потому что мы не видим соответствующих показателей, и просодического отличия повествовательных предложений от императивов, оптативов и вокативов, если судить также и по другим молитвам (это молитвы Утреннего цикла, из которого взято данное чтение «Отче наш») нет. Возникает вопрос, какие различия здесь есть. Как распределяются подъемы тона и на каких словоформах в строке они фиксируются? Вопрос о членении речи, в частности литургической, на высказывания, строки и законченные тексты мы здесь оставляем в стороне. Будем считать, что тексты расчленены: в обычном стиле речи — на предложения, в литургическом и поэтическом — на строки в соответствии с письменной формой соответствующих текстов.

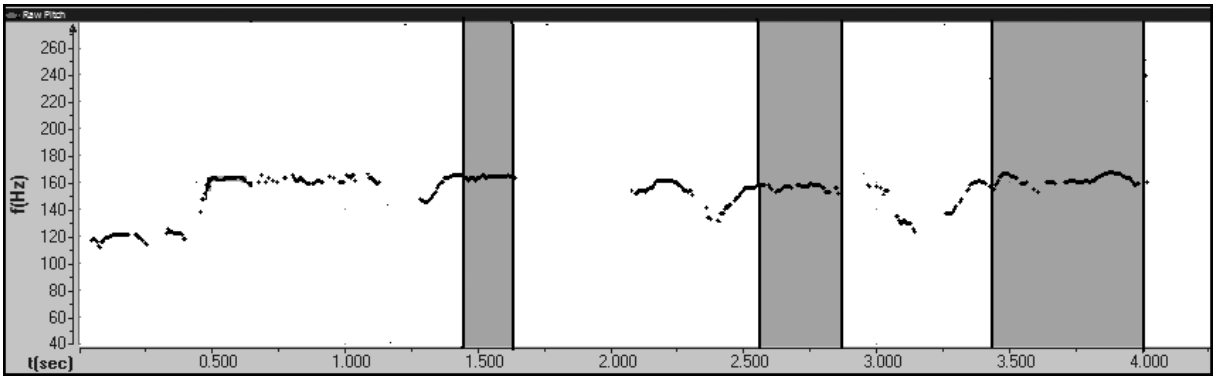
Какие конкретные просодические признаки литургического чтения здесь выделяются? Анализ показывает, что это три параметра: изменения частоты основного тона, длительность и выбор акцентоносителя. Наиболее частотная реализация первого признака — это подъем на ударном слоге словоформы-акцентоносителя плюс ровные заударные. Такой акцент наблюдается на словоформах будетъ и хлѣб в примере (1). Такой тип изменения тона последовательно проходит через всю молитву и другие молитвы цикла.

Вторая реализация признака изменения частоты основного тона — тот же подъем с ровными заударными, но в меньшем диапазоне частот. В примере (1) эта модель реализуется на словоформе *воля*. Подъем с ровными заударными в сокращенном диапазоне частот тоже проходит через все молитвы цикла.

Реализация признака длительности — это тот же подъем с ровными заударными плюс продленная длительность. Этот акцент возникает не в начале, а в конце строки, как в примере (3) на словоформе *лукаваго*.

(96) *и не введи насъ во искушенїе, но избави насъ от лукаваго.*

Тонограмма 3



и не введи *насъ* *во искушенЇЕ, но* *избаВИ НАСЪ* *от* *лукаВАГО*

180 мсек 260 мсек 550 мсек

(VIII) (IX)

Тонограмма 3 демонстрирует длительность последовательности двух заударных слогов в разных позициях в строке: длительность заударных в конце строки здесь в 2–2,5 раза больше, чем в других позициях во фразе и во всем тексте в целом. Этот акцент мы интерпретируем как маркер конца, но не предложения, а текста. На словоформе *лукаваго* молитва кончается. Продленную длительность мы находим и в других молитвах цикла, оканчивающихся на *Аминь!* В конечных строках молитв, таким образом, наблюдается не один акцент, а, как минимум, два: начальный и конечный. В примере (3) конечная строка (IX) имеет два подъема — на начальной словоформе *избави* и на конечной — *лукаваго*, причем словоформа *лукаваго* имеет сверхдолгие заударные.

Последний релевантный признак — выбор словоформы-акцентоносителя. В молитве акцентоносителем подъема в максимальном для строки диапазоне служит первое фонетическое слово. Этот акцентоноситель выбирается чисто механически — в обход принципов выбора акцентоносителя, существующих в языке и основанных на синтаксических приоритетах и других факторах. В примере (1) это начальные словоформы двух строк *будетъ* и *хлѣбъ*, а в примере (3) — *введи* и *избави*.

Далее. Носителем подъема в меньших диапазонах частот служит акцентоноситель, который выбирается по синтаксическим правилам и который мы называем дефолтным. Он был бы в предложении, если бы его выбор не механическим (т.е. чисто линейным), а собственно языковым (или синтаксическим). В примере (1) *да будетъ волѧ Твоѧ* (строка (IV)) дефолтный акцентоноситель — словоформа *волѧ*, а в строке (V) *хлѣбъ нашъ насущныи даждь намъ*

днесь это словоформа *хлѣбъ*, которая одновременно служит не только дефолтным акцентоносителем, но и первым фонетическим словом.

Не останавливаясь здесь на языковых принципах выбора акцентоносителя, для объяснения результата такого выбора мы предлагаем следующий эмпирический прием. Чтобы выбрать акцентоноситель в церковнославянском предложении, нужно взять соответствующее русское предложение с нейтральным порядком слов[3]. Так, для строки (V) *хлѣбъ нашъ насущныи даждь намъ днесь* в примере (1) это предложение (4):

(97) *Дай нам сегодня наш насущный хлеб.*

Далее. В этом предложении выделяется конечная словоформа, потому что в нейтральных предложениях акцентоноситель располагается в абсолютном исходе. В русском предложении (4) это словоформа *хлеб*, а в предложении (5), соответствующем строке (V), это словоформа *воля*:

(98) *Да будет твоя воля.*

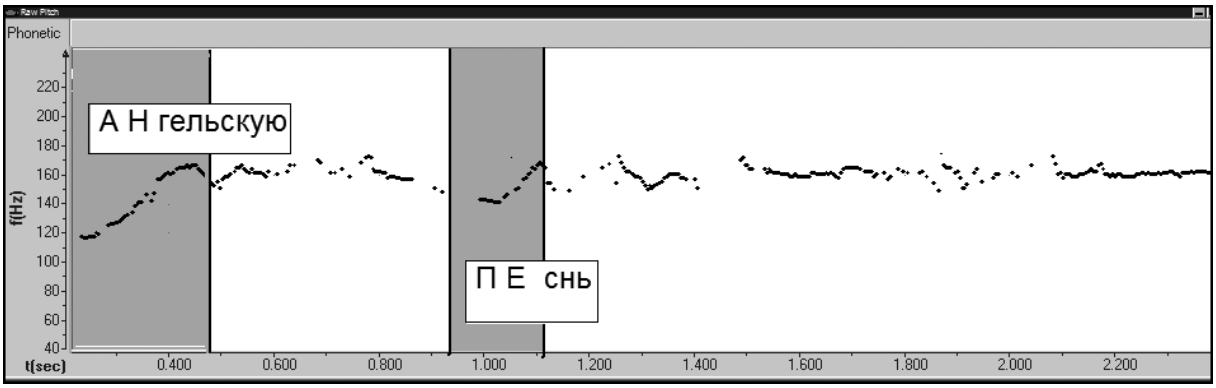
[3] Поясним, что под нейтральным порядком мы понимаем такой, у которого вклад линейной структуры в семантическую минимальный.

Эта словоформа предсказывает, каков будет носитель дефолтного акцента в церковнославянском предложении. Доказательство того, что дефолтные акцентоносители русского языка совпадают с носителями второстепенных — меньших по диапазону — подъемов в тексте православных молитв, если они находятся не в начале строки, мы здесь опускаем. Оно следует из принципов выбора акцентоносителя, о которых говорилось выше (см. Янко 2001: 68–84; 2008: 38–60), а также из гипотезы о том, что принципы выбора акцентоносителя в русском и церковнославянском языках совпадают. Некоторые несовпадения выбора акцентоносителя для разных языков изложены в (Янко 2008: 73–82; 274–282), но эти расхождения не затрагивают тех случаев, которые рассматриваются здесь.

В примере (6) из Тропарей Троичных мы наблюдаем два акцентоносителя: начальный — это словоформа *ангельскую* — и дефолтный, который отмечен более слабым акцентом, — это словоформа *песнь*, ср. русское предложение (7) с конечным акцентоносителем *песнь*. На тонограмме (4) области, соответствующие ударным слогам словоформ-акцентоносителей *ангельскую* и *песнь*, выделены курсорами и заливкой.

(99) *…и ангельскую песнь вопием Ти, Сильне…*

Тонограмма 4



и а н г е л ь с к у ю п е с н ь в о п и е м Т и, С и л ь н е

(100) *Мы поем тебе ангельскую песнь.*

В примерах (1) (строка (IV)) и (6) начальная словоформа не совпадает с дефолтным акцентоносителем, поэтому в этих строках по два акцентоносителя — начальный и дефолтный.

На тонограмме 5 даны записи двух чтений одного и того же фрагмента (8) из Молитвы ко Пресвятой Троице: верхняя панель отражает речь того же священника, речь которого была представлена примерами (1)–(7) выше, а нижняя панель отражает чтение другого священника. Сравнение этих двух чтений (а также других молитв, прочитанных этими священниками) говорит о том, что признаки литургического чтения, выделенные выше, скорее всего, не являются случайными, так как совпадений в артикуляции достаточно много.

(101) *…ниже погубил мя еси со беззаконьми моими; но человеколюбствовал еси обычно…*
‘и не погубил меня со беззакониями моими,
но как всегда явил Свое человеколюбие’

Пример (8) с частотными показателями начала строки на словоформах *ниже́* (со значением ‘даже не’; ударение на втором слоге) и *но* (союзе), служит

дополнительным соображением в пользу того, что акцентоносители в данном типе чтения выбираются по линейному принципу, потому что других объяснений просодическому выделению неполнозначных словоформ *нижé* и *но* мы не видим. Пример (8) оба священника произносят с повышением частоты основного тона на начальных словоформах строк. В этом — основном выделяемом нами признаке православного литургического чтения — они единодушны. Кроме того, темп обоих чтений совпадает: оба фрагмента занимают практически одинаковое время звучания. Различия в чтении есть, но, как кажется, они не имеют принципиального характера. Чтение, представленное на верхней панели, содержит дефолтный подъем частоты на *обычно* и небольшой подъем на ударном слоге словоформы *человеколюбствовал*, а чтение, представленное на нижней панели, более монотонно. Другая характерная черта чтения второго священника состоит в том, что у него более существенно, чем у первого, удлиняются конечные словоформы строк, причем не только тех, которые служат концом молитв.

Тонограмма 5

Итак, в литургическом чтении мы предлагаем выделять три типа подъема — 1) подъем в большом диапазоне частот, 2) подъем в усеченном диапазоне частот и 3) подъем в условиях продленной длительности. Подъем в большом диапазоне частот маркирует начало строки. Подъем в усеченном диапазоне частот — факультативно — маркирует дефолтный акцентоноситель соответствующего предложения, если он находится не в начале строки. Этот акцентоноситель отражает синтаксическую структуру предложения. И наконец, подъем со сверхдолгими заударными формирует конец текста. Соответственно, в конечной строке, как минимум, два акцентоносителя: первый формирует строку, второй — завершает текст. Показателей других языковых значений мы в данном типе чтения не видим.

2. «Pater noster» vs. «Отче наш»

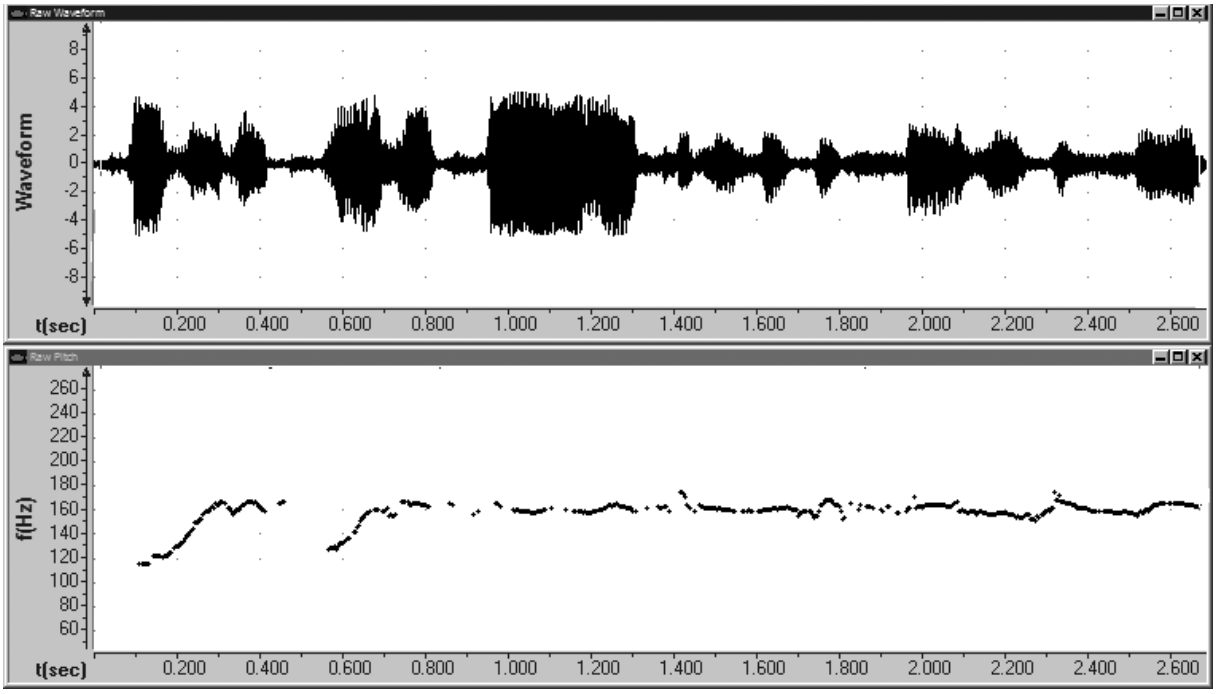
Обратимся к историческому вознесению молитвы Pater noster папой Пием XII при вошествии на папский престол. Чтение папы выдержано в торжественном стиле, молитва читается прочувствованно с энергичным выражением всех иллокутивных сил. Так читают актеры стихотворные произведения. Обратимся к тексту молитвы.

Pater noster, qui es in caelis,
sanctificetur nomen tuum,
adveniat regnum tuum,
fiat voluntas tua,
sicut in caelo, et in terra.
Panem nostrum quotidianum da nobis hodie;
et dimitte nobis debita nostra,
sicut et nos dimittimus debitoribus nostris;
et ne inducas nos in tentationem,
sed libera nos a Malo. Amen.

Сравним православное литургическое чтение (пример (9), строка (IV), тонограмма 6) и чтение соответствующего фрагмента молитвы «Pater noster» Пием XII (пример (10), тонограмма 7). (Поясним, что, кроме тонограмм, здесь — на верхних панелях — приводятся и осцилограммы, отражающие, в частности, структуру пауз, которые на осцилограмме выглядят как отрезки с минимальной амплитудой колебаний).

(102) *да будетъ волѧ Твоѧ, яко на небеси и на земли.*

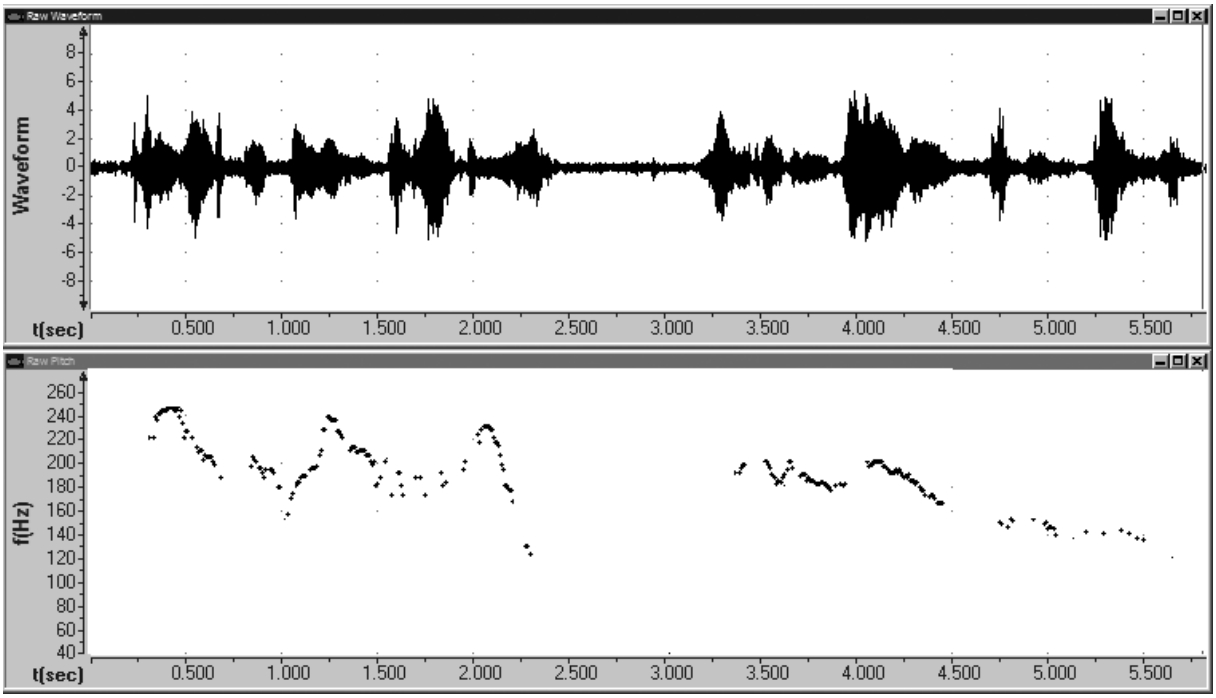
Тонограмма 6



да будетъ волѧ Твоѧ, яко на небеси и на земли

(103) *...fiat voluntas tua, sicut in caelo, et in terra.*

Тонограмма 7

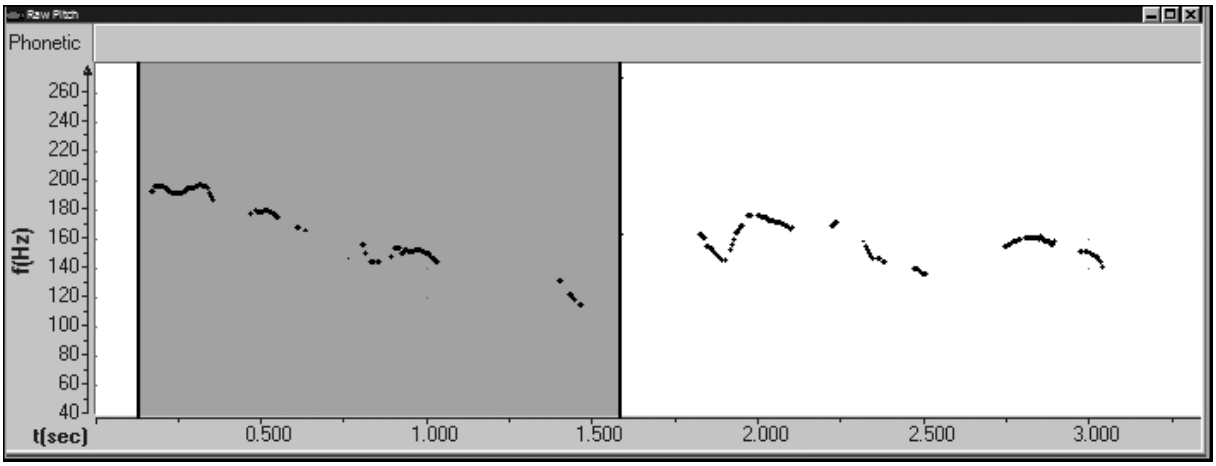


fiat voluntas t u a, sicut in caelo, et in terra.

Фрагмент *Fiat voluntas tua* читается в полном соответствии с просодическими средствами выражения оптатива — конкретный язык в данном случае роли не играет — с высоким началом на *fiat* и рельефным падением на *tua*. Между главным предложением и придаточным с *sicut* имеется полноценная прочувствованная пауза. В исходе всего предложения частота падает, что характерно для данного типа речевых актов в речи с просодически ясными (в данном случае — даже подчеркнутыми) иллокутивными силами. Выбор акцентоносителей находится в полном соотвествии с принипами выбора акцентоносителя в языке: об этом говорит и высокая частота на ударном слоге словоформы *fiat*, и крутое падение на *tua*, высокий тон на ударном слоге словоформы *caelo* — начале союзной группы, состоящей из двух связанных союзом *et* конъюнктивных членов, и падение на ударном слоге завершающей словоформы *terra*, которое продолжается на заударном слоге. Как мы видим, в лаконичном православном чтении такой игры иллокутивных просодий нет. Существенным отличием чтений служит темп: литургическое чтение фрагмента вдвое короче, чем чтение папы. Приведем другие примеры из «Pater noster». Обратимся к начальным словам (тонограмма 8):

(104) *Pater noster, qui es in caelis,*

Тонограмма 8

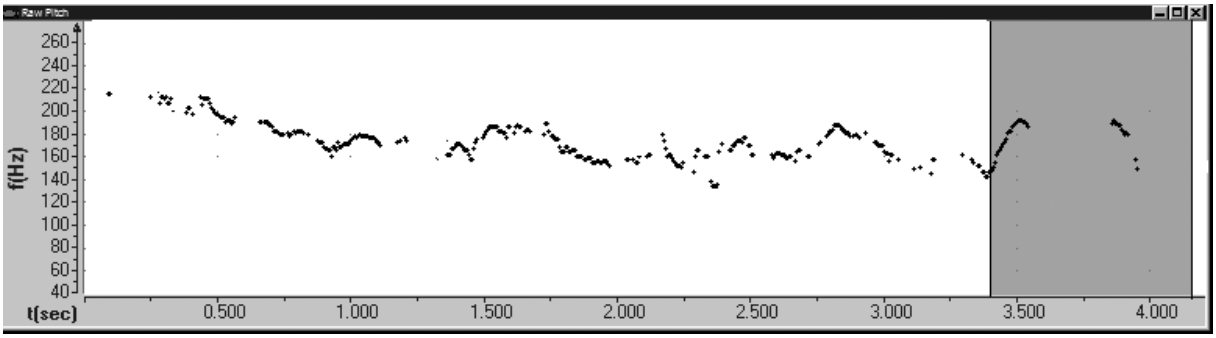


P a t e r n o s t e r, q u i es in c a e l i s

Вокатив *Pater noster* начинается с высокой частоты на первом ударном слоге словоформы *pater*, на второй словоформе вокатива частота падает, между вокативом и определением к нему *qui es in caelis* имеется существенная пауза. Словоформа *caelis* маркирует незавершенность текста — на ней фиксируется подъем на первом ударном слоге и падение на заударном. Опять же и иллокутивное значение, и незавершенность текста выражены весьма рельефно. Показатель незавершенности с подъемом на ударном слоге и падением на заударных мы находим и в примере (12). Акцентоноситель незавершенности — словоформа *nostris*. Это основной — немаркированный, т.е. не отягощенный другими сопутствующими незавершенности значениями, показатель связности текста во многих языках. Для этого показателя характерно падение на заударных, если они есть. Этот показатель частотен в обычной речи, а для православной литургической традиции он нехарактерен.

(105) *...sicut et nos dimittimus debitoribus nostris ...*

Тонограмма 9



...sicut et nos d i m i t t i m u s d e b i t o r i b u s n o s t r i s ...

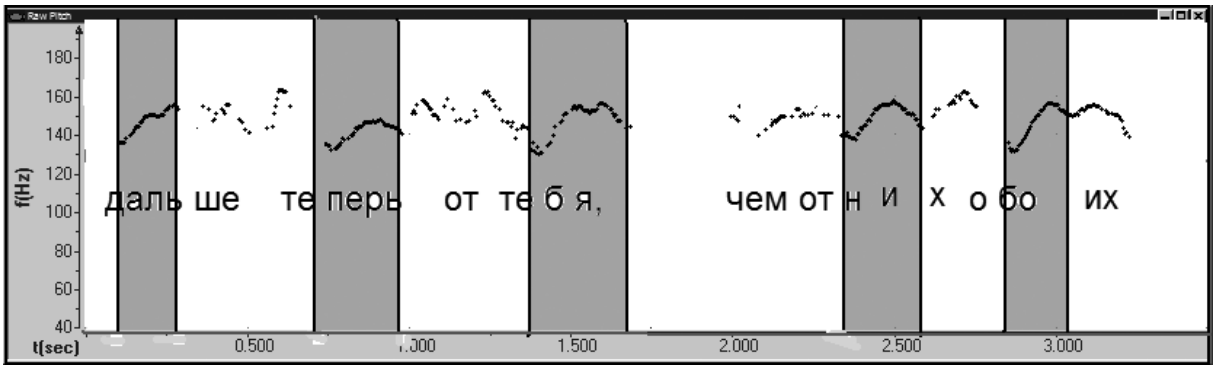
Итак, папа читает молитву с сохранением всех интонационных показателей иллокутивных и других коммуникативных значений, воспроизводя свое общение с Всевышним, как общение с человеком. Он взывает к Богу, не цитируя Иисуса Христа и не ведя за собой паству, а от своего собственного имени. Это отличает его чтение от православного литургического речитатива, который отмечен снятием стандартных иллокутивных сил.

3. «Ниоткуда с любовью» И. Бродского

Обратимся к чтению Иосифом Бродским одного из его стихотворений. Это чтение представляется нам достаточно характерным для стилистики Бродского. Тонограмма 10 выбранного наугад фрагмента (13) из чтения стихотворения «Ниоткуда с любовью» демонстрирует последовательность подъемов продленной длительности на каждом фонетическом слове текста с относительно ровными или вибрирующими заударными. Это чтение без начал и концов строк, без разделения высказываний на сообщения и вокативы:

(106) *...дальше теперь от тебя, чем от них обоих...*

Тонограмма 10



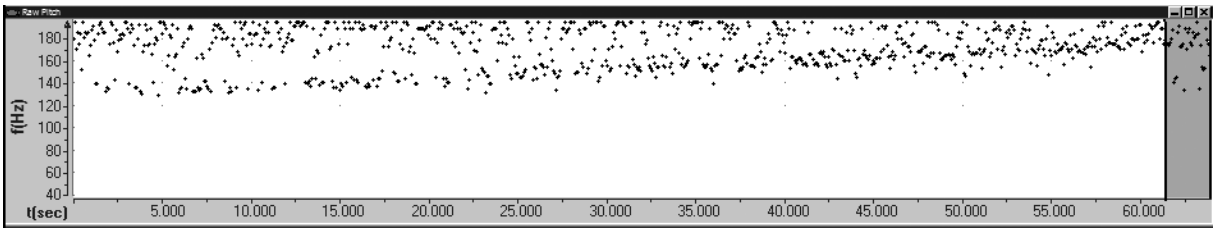
На ударных слогах словоформ *дальше*, *теперь*, *тебя*, *них* и *обоих* в данном отрезке мы видим подъемы в сокращенных диапазонах частот и заударные продленной длительности. Ударные слоги на тонограмме выделены курсорами и заливкой. Единственное формальное значение, которое удается приписать данной просодии, если отвлечься от поэтической сверхзадачи автора, это членение текста на ритмические группы — в данном случае — на полнозначные слова, включая местоимения. Все чтение достаточно последовательно выдержано в подобной манере. Между тем, если перейти от фрагмента к тексту в целом, можно обратить внимание и на другие функции просодии, которые наблюдаются при такой стилистике чтения. На тонограмме 11 представлен сжатый график показателей частот всего стихотворения в целом.

Мы видим, что к концу стихотворения средняя частота артикуляции последовательно повышается, что говорит о растущем эмоциональном напряжении поэта. Минимальные частоты в начале чтения — около 120 герц, в конце — около 160. Это не средство выражения языковых значений и не специальный способ представления душевного накала, а отражение физиологического состояния человека в зеркале частот тона его голоса. При эмоциональном возбуждении частота голоса повышается. И только два последних фонетических слова несут на себе маркер конца текста, что выражается в видимом снижении частоты. Соответствующая область на тонограмме 11 выделена курсорами и заливкой. Это словоформы *зеркало* и *повторяя*. Ср. кривую частоты финального фрагмента (12) на тонограмме 12.

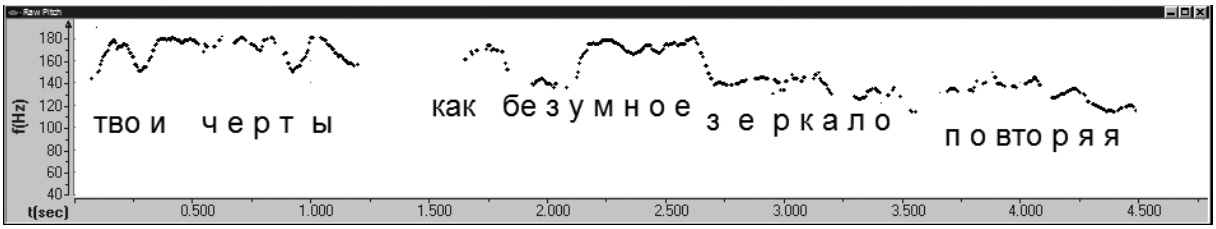
(107) *…твои черты, как безумное зеркало, повторяя.*

Фрагмент *твои черты, как безумное* артикулируется на высоких частотах, а на *зеркало* и *повторяя* происходит общее снижение частоты на ударных с сохранением вибрирующего рисунка на заударных. Стихотворение кончается, перед на ми маркер конца текста — возвращение к частотам начала. Автор читает стихотворение, подчеркнуто игнорируя концы строк не только с помощью просодии, но, прежде всего, с широким использованием синтаксического средства отказа от членения строфы на строки — приема enjambement. Паузы не несут семантической нагрузки, они используются автором только для восстановления дыхания. При подобном чтении различия между текстом и предложением стираются.

Тонограмма 11



Тонограмма 12



4. Аят аль Курси

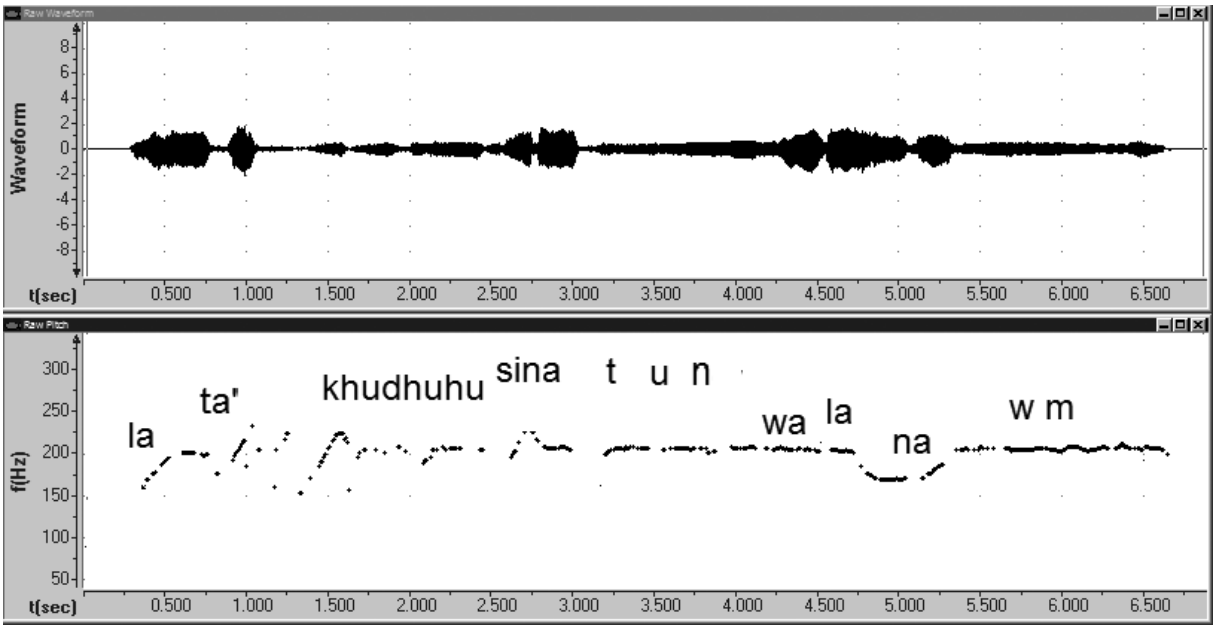
В данном подразделе сделана попытка обнаружить элементы формального просодического членения мусульманской молитвы. Перед нами одно из немногих исполнений этой молитвы, которое ближе к речитативу, чем к пению. Другие доступные нам записи этой молитвы представляют собой пение, которое нуждается в других методах анализа. Читает шейх Аль-Хусейни аль-Азази. Обратимся к тексту молитвы в латинской транслитерации.

Ayat al Kursi
Bismillahir-Rahmanir-Raheem
'Во имя Аллаха, милостивого и милосердного'
Allahu la ilaha illa Huwa, Al-Haiyul-Qaiyum
'Аллах — это тот, кроме которого, нет божества. Он живой, вечно существующий'
La ta'khudhuhu sinatun wa la nawm,
'не одолевают его ни дремота, ни сон.'
lahu ma fis-samawati wa ma fil-'ard
'Ему принадлежит все, что в небесах, и все, что на земле,'
Man dhal-ladhi yashfa'u 'indahu illa bi-idhnihi
'кто перед ним заступится, без его разрешения?'
Ya'lamu ma baina aidihim wa ma khalfahum,
'Он знает, что было перед ними и знает, что будет после них,'
wa la yuhituna bi shai'im-min 'ilmihi illa bima sha'a
'они овладевают из Его знаний только тем, что он пожелает.'
Wasi'a kursiyuhus-samawati wal ard,
'Трон Его объемлет Небеса и Землю'
wa la ya'uduhu hifdhuhuma Wa Huwal 'Aliyul-Adheem
'и не тяготит Его охрана их, истинно.'

Это чтение характеризуется двумя релевантными подъемами продленной длительности — в начале и в конце строки. В примере (15) релевантные движения тона фиксируются на слоге [la] в начале строки и на [nawm] (плюс сверхпродленная долгота) — в исходе строки, ср. тонограмму (и осциллограмму) 13.

(108) *La ta'khudhuhu sinatun wa la nawm*
'не одолевают его ни дремота, ни сон'

Тонограмма 13

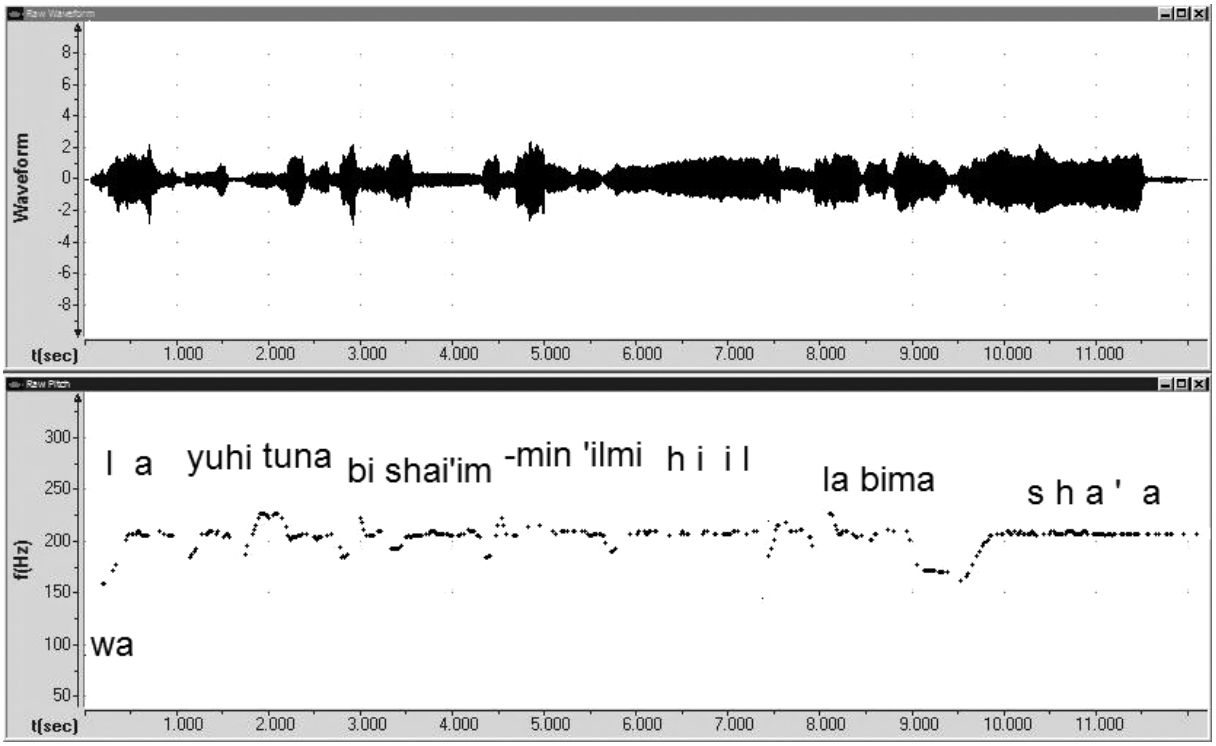


И аналогичный пример с начальным подъемом на [la] и конечным — на [sha/a]:

(109) *wa la yuhituna bi shai'im-min 'ilmihi illa bima sha'a*
'они овладевают из Его знаний только тем, что он пожелает'

В отличие от других рассмотренных здесь текстов структура чтения Аль-Азази аль-Хусейни представлена в несколько упрощенном виде, однако основная тенденция к выделению начал и концов строк, которая хорошо воспринимается на слух, в нем выдерживается. Конец текста просодически специально никак не отмечен.

Тонограмма 14



5. Схемы строк

Мы рассмотрели четыре типа просодической артикуляции предложений и/или строк. 1) Первый тип характеризует речевые акты-невопросы с подъемом в начале предложения и падением в конце. Срединная артикуляция может быть различна — с высокими и низким заударными в зависимости от выражаемых значений. Эта схема отражает обычную речь, а из рассмотренных выше примеров — обращение православного священника к пастве и чтение папой Пием XII молитвы “Pater noster”. Мы считаем чтение папы разновидностью обычной речи в ее подчеркнуто рельефном просодическом исполнении: диапазоны частот и паузы здесь существенно увеличены по сравнению с показателями обыденной неподготовленной речи. В беседе священника с паствой иллокутивные силы и связность текста тоже выражены просодически весьма отчетливо. Оба типа чтения характеризуются значениями различительных просодических признаков, близкими к максимальным. 2) Православное литургическое чтение характеризуется подъемом тона с ровными заударными в начале каждой неконечной строки текста и дополнительным подъемом плюс увеличенная длительность — в самом конце текста. Кроме того, в строках (неконечных и конечных) может — факультативно — быть дополнительный подъем в суженном диапазоне частот на акцентоносителе, который отражает синтаксическое членение. 3) Иосиф Бродский отмечает подъемами с ровными (или качающимися) заударными каждое фонетическое слово текста плюс падение частоты в абсолютном конце текста по просодическому типу ремы. 4) Аят аль Курси читается шейхом Аль-Азази Аль-Хусейни с двумя подъемами в начале и в конце каждой строки.

***

В данной работе было рассмотрено несколько образцов чтения молитв и поэтических произведений, пилотный анализ которых позволяет сделать некоторые предварительные выводы о просодической структуре предложения в некоторых традициях. Просодия может не соответствовать иллокутивным силам, которые предполагаются сегментными средствами соответствующих высказываний. К таким традициям можно отнести православную литургическую традицию, чтение поэтами их собственных стихов, а также мусульманскую традицию чтения молитв, лежащую на границе с пением. Единицы просодии при чтении со «снятой» иллокутивностью, имеют, тем не менее, функцию членения текста, но не на коммуникативные компоненты предложений, а на другие единицы: при православном литургическом чтении отмечается начало строки и конец текста, в мусульманской молитве — начало и конец строки, а при чтении Бродским его собственного стихотворения отмечается каждое фонетическое слово текста и конец текста. Отказ от выражения иллокутивных значений имеет различные — высшие — цели, которые могут служить объектом анализа у соответствующих специалистов.

Литература

1. *Владышевская Т. Ф.* К вопросу о роли византийских и национальных руссских элементов в процессе возникновения древнерусского церковного пения // Материалы IX Международного съезда славистов в Киеве. М., 1983, С. 10–31.
2. *Кутузов Б.* Экфонетика в православном богослужении // Журнал Московской патриархии. № 7. 1999.
3. *Пантелеева И. А.* Влияние античной декламации на формирование музыкальных основ религиозного дискурса // Наука. Релігія. Суспільство», № 1, 2009.
4. *Прохватилова О. А.* Православная проповедь и молитва как феномен современной звучащей речи. Волгогр. гос. ун-т., Волгоград, 1999. 362 с.
5. *Русская грамматика* — Т. 1, М.: Наука, 1982.
6. *Янко Т. Е.* Коммуникативные стратегии русской речи. М.: ЯСК, 2001.
7. *Янко Т. Е.* Интонационные стратегии руссской речи в сопоставительном аспекте. М.: ЯСК, 2008.

Восприятие темпа речи и некоторые находки в сфере моделирования речевой ритмической структуры эстоноязычной речи

Speech rate perception and some findings of modelling speech rhythmicity In estonian

Meelis Mihkla (meelis@eki.ee), **Indrek Hein** (kiisu@eki.ee),
Mari-Liis Kalvik (mariliis@eki.ee), **Indrek Kiissel** (indrek@eki.ee)
Institute of the Estonian Language

Статья посвящена проблемам восприятия различного темпа речи слепыми и зрячими, а также вопросам моделирования ритмики речи. Выяснилось, что «тренированные слепые» предпочитают гораздо более быстрый темп речи, чем зрячие. При обращении к трем степеням долготы в эстонском языке оказалось, что соотношение долготы гласных в ударном и безударном слоге является важнейшим признаком определения степени долготы.

1. Introduction

The necessity for the two relatively different studies arose in the course of developing an audio system enabling the blind to listen to reference texts and audio books. By means of the on-line system of the Estonian Library for the Blind http://www.epr.ee/kalev the visually impaired can have texts (news, newspapers, magazines and books) read for them and listen to audio books over the Internet. The use of the system revealed that many blind people wish to hear the news and newspaper articles at a considerably higher speech rate than normal. As the system is server-based it cannot afford users tuning the rate smoothly as it would make the system too cumbersome and slow. Hence the need to find some optimal rates to supplement the user menu with two speech rates from the quicker-than-normal range, say, quick and very quick.

Perception of rapid rate of speech and the limits of its temporal compression have been discussed in several studies (Asakawa a. o. 2003 and Moos, Trouvain 2007). It has been found that very rapid speech is preferred and perfectly understood by trained people, i. e. those with an everyday experience of a screen reader and a speech synthesizer. In the course of a joint study by the Universities of Saarland and Tübingen (Moos a. o. 2008) the brains of blind and of sighted subjects were scanned while they were listening to rapid speech; it was found that in the blind a very quick rate was systematically accompanied by activation in the brain zone that in the sighted is used for processing visual information.

Recently the mechanisms of speech rhythmicity have been drawing increasing attention. The focus lies on different aspects of the vowel onset of the stressed syllable (Keller, Port 2007), which play a decisive role in enhancing the naturalness of synthetic speech (Keller 2007). The temporal structure and rhythmicity of speech is partiularly important in Estonian, where foot is the arena for the phonological opposition of three quantity degrees (Q1, Q2, Q3) to realize. In principle, Estonian quantity degrees are defined in the same spirit as the newer approaches to speech rhythmicity (for an overview see Keller, Port 2007): in both the fulcrum is the onset of the rime of a stressed syllable. Quantity degrees are, in essence, suprasegmentals (Lehiste 1997), whose definition, on the acoustic level, relies on the durational relations of the rime of the stressed syllable and the nucleus of the unstressed syllable, plus the F0 contour (Ross, Lehiste 2001), making up a complementary system. In addition, the durational relations of consecutive phones and some other features have been suggested as important (Eek, Meister 2003). The present study compares different parameters of quantity and, by statistical modelling, evaluates their significance.

2. Speech rate perception

Although an on-line audio system is meant for the visually impaired mainly, our test of speech rate perception was also applied to a group of sighted subjects, for comparison. This was meant to answer such questions as: What speech rates are preferred by the blind vs. the sighted? Is the preference of very quick speech rate by the visually impaired a myth or not? Is there such a thing as an optimal speech rate?

2.1. 1 Subjects

The test was taken by 58 blind or heavily visually impaired subjects (29 female and 29 male, aged 14–79) and by 56 sighted subjects (41 female and 15 male, aged 18–58). For all subjects, Estonian was the mother tongue.

2.2. Test material

The stimuli for the test of speech rate perception were generated from two audio books ("American tragedy" by T. Dreiser, male voice, and "Das erste Mal und mehr" by E. Stein-Fischer, female voice, both in Estonian translation) and some news fragments, synthetic voice. The latter was produced by a diphone-based Estonian text-to-speech synthesizer (Mihkla, Meister 2002), using an MBROLA synthesis motor. The synthetic voice was generated in two variants, one using a rule-based prosody model (SYNT1) the other a statistical one (SYNT2). This was to test the impression of the blind that in the case of a synthesizer with a statistical prosody module, rate quickening would lower the quality of output speech.

For the female voice the natural reading rate was 135 words per minute, the male voice making 122 words per minute; the synthesizers were tuned to match the female rate. For each voice, eight speech samples of 35–55 sec were generated, each of a different speech rate (see Fig. 1).

81	108	**135**	162	189	216	243	270	words/min
60	80	**100**	120	140	160	180	200	%

Figure 1. Speech samples as stimuli of different speech rates (natural speech rate 100 = 135 s/min).

The limits of temporal compression had been first agreed upon with some "trained persons", who have everyday practice of listening to synthetic speech. In their opinion the maximal rate for listening to prose texts is twice the usual rate. During test preparation a few older representatives of the visually impaired suggested that they might perhaps wish to listen to some paragraphs at a slower pace. Thus we added two slower samples (0.8 and 0.6 times the natural rate). The rate of the samples from audio books was regulated by means of the signal processing program Adobe Audition 3 for high precision time compression with time stretch (preserves pitch).

The subjects were exposed to the speech rate stimuli by voice series presented in a random order. The appropriateness of the speech rate was asked to be evaluated in a five-point system (5 — the best, 4 — good, 3 — tolerable, 2 — uncomfortable, 1 — unsuitable , i. e. unintelligible, too quick or too slow)

2.3. Results

Fig. 2 presents the average blind vs. sighted scores for different speech rates. The left diagram shows the scores given for the female voice, the right one for the synthetic voice SYNT1. According to the diagrams the blind prefer the speech rates 1.2 and 1.4, which are 20 % and 40 % quicker than natural, respectively. The sighted think highly of natural speed, but 1.2 and 1.4 are not considered bad either.

The following figure (3) demonstrates the points scored by the male and the synthetic voice SYNT2 (the left and right diagrams, respectively). For a male voice 1.2 was considered the best speech rate both by the sighted and the blind. The results are probably due to the natural speech rate for a male voice, which is about 10 % slower than the rest (122 s/min *versus* 135 s/min).

Although the subjects were asked to evaluate the suitability of the speech rate only, not the pleasantness of the voice, synthetic speech scored almost a point lower, on average, than human speech. However, the prosody module of the synthesizer (SYNT1 vs. SYNT2) does not seem to have influenced the score significantly at quicker speech rates. Thus the results fail to support the idea that the quality of the SYNT2 voice might deteriorate at quicker rates.

The test proved that the ratings of the blind and the sighted differ far less than first believed. Figures 2 and 3 present the average scores from all visually impaired subjects. However, the blind include many who seldom use a computer, if at all, and who thus lack the experience of listening to synthetic speech at different speech rates. Figure 4 presents the average scores given to different speech rates vs. the previous experience, in years, of the visually impaired subjects with a computer and screen reader. The results reveal an obvious tendency that in the visually impaired, longer practical "training", i. e. experience of using the above devices causes an increase in the ability of understanding rapid speech.

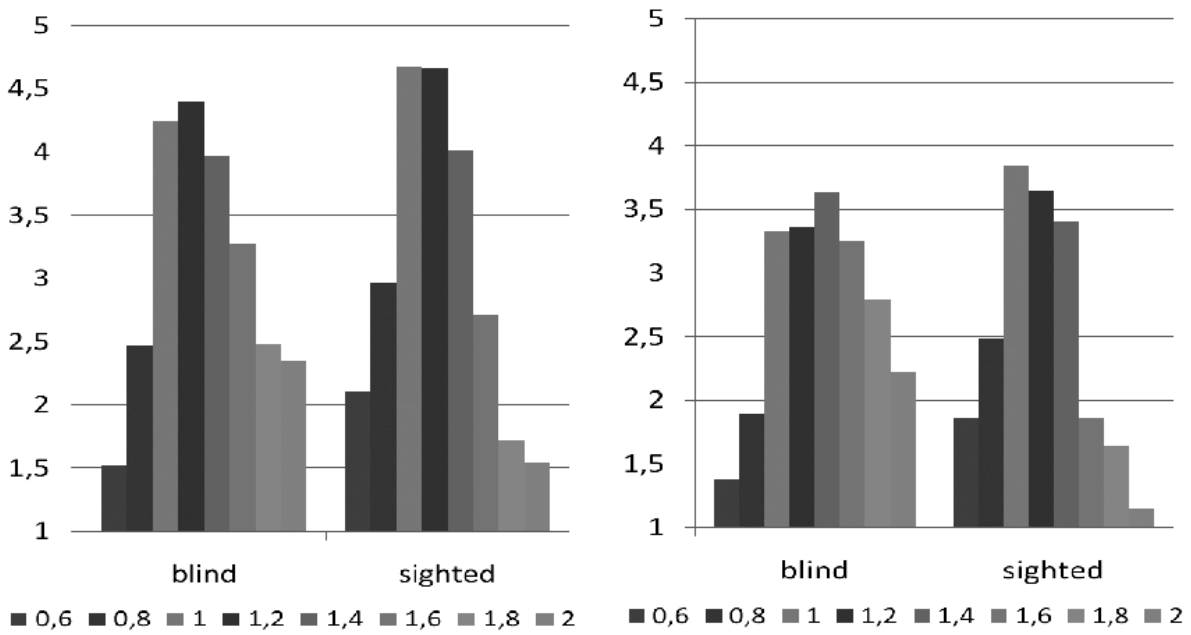


Figure 2. The average points from the blind and the sighted for the female and the synthetic (SYNT1) voice

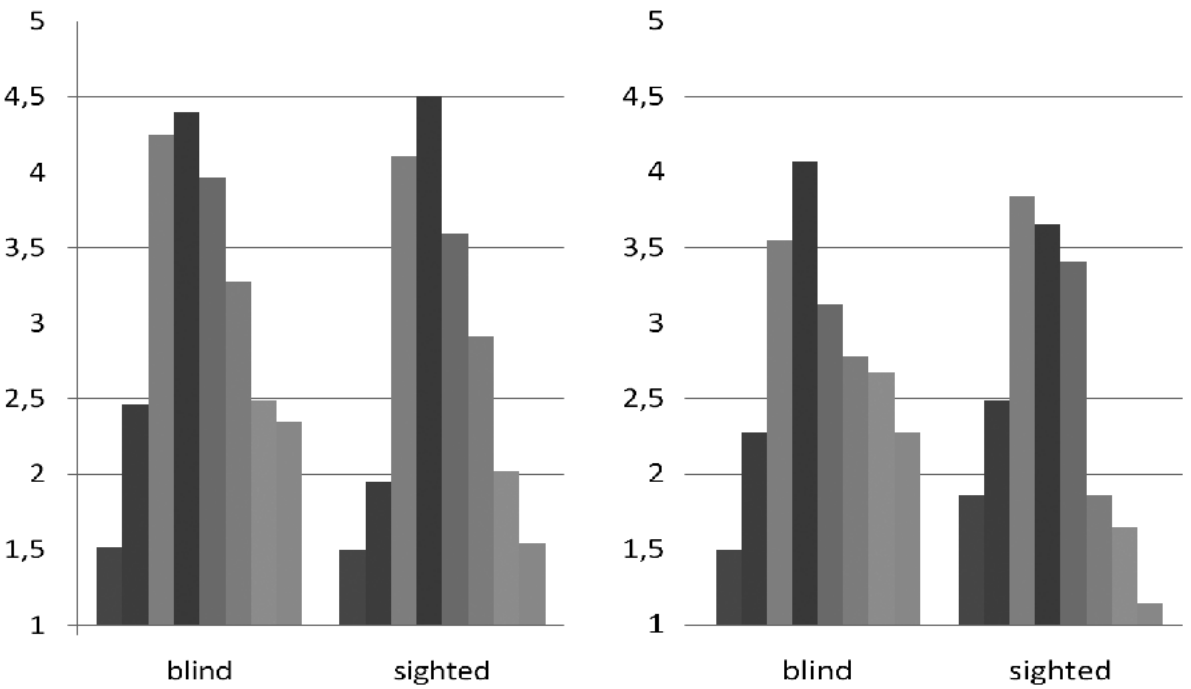


Figure 3. Average points from the blind and the sighted for the male and the synthetic (SYNT2) voice

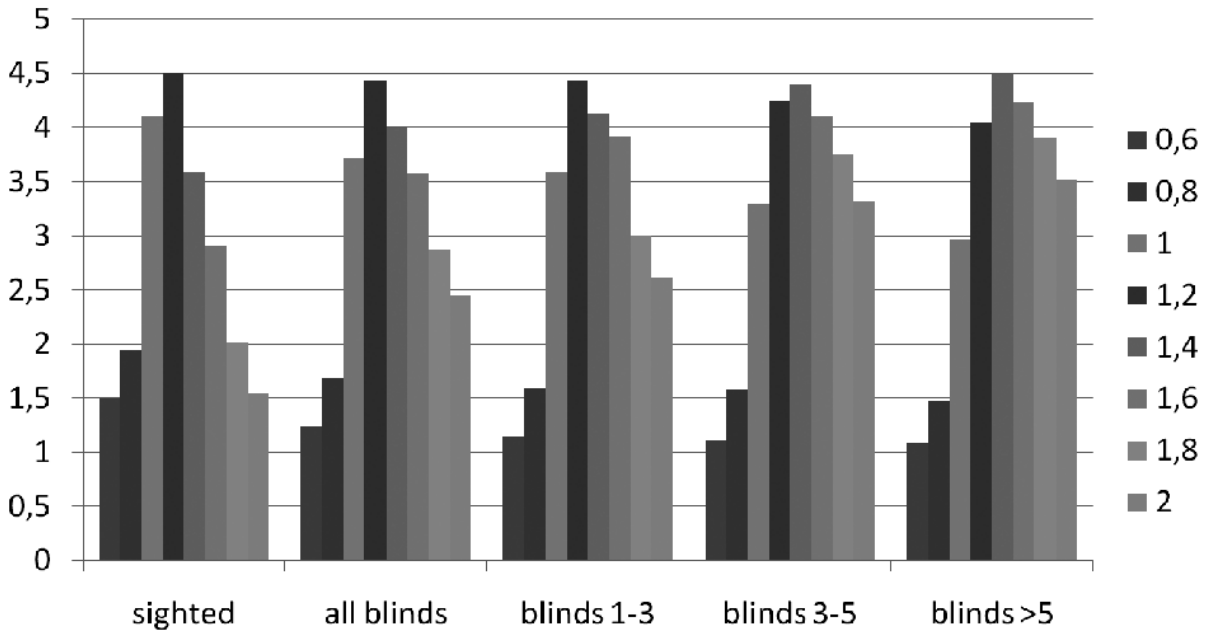


Figure 4. The average scores given to different speech rates vs. the previous experience, in years, of the visually impaired subjects with a computer and screen reader.

3. Recognition and modelling of speech rhythmicity

3.1. Introduction

For quite a while now the problem of the distinctive features of the three quantity degrees — short, long and overlong (Q1, Q2, Q3) — observed in standard Estonian have been subject to discussion among phoneticians. Up to the present the duration ratio between the stressed and unstressed syllable in a foot and, in particular for Q2 and Q3, a difference in their pitch curves have been considered the most important parameters to describe and analyse the quantity degrees of words from differently structured speech, lab-generated as well as spontaneous. As a result of several studies it has been found that the duration ratio between the first and second syllables is 2:3 for Q1, 3:2 for Q2, and 2:1 for Q3 (Lehiste 1960, Ross, Lehiste 2001).

The present study investigates, in addition to the traditional duration ratios presented, the ratio of adjacent segments and discusses the possible role of intensity. Manifestation and perception of phonetic quantity make up a complicated system, where different conditions may evoke different combinations and different salience of parameters. In this study we seek additional parameters possibly depending on phonetic quantity, by weighing their relevance with statistical methods. As our aim in modelling speech temporal structure is high quality synthetic speech we need the best possible parameters to describe and discriminate the three Estonian degrees of phonetic quantity, so that each degree could get its own model.

In the present study we test some of the ideas suggested by Arvo Eek. In the first place, Eek and Meister have, on the basis of perception tests, created a theory focused on adjacent phones within the main stress syllable and the successive syllable. In a two-syllable word with a vowel-centred structure CV(::)CV, the duration ratio of the vowel (V1) to the consonant (C1) of the first, stressed syllable is supposed to discriminate Q1 words, which have a short V1, from Q2 and Q3 words, whose V1 is long. Q3 words can purportedly be distinguished from Q1 and Q2 words by the duration ratio of the vowel (V2) to the inter-vowel consonant (C2) of the second, unstressed syllable. Starting from a belief that the perception of duration difference between two adjacent phones is not possible unless one is 20–25 % longer than the other they have calculated 1.4 as the limit of duration difference. Words have a short V1 (thus qualifying as Q1 words) if their V1:C1 ratio is less than 1.4, while for a long V1 (signalling of Q2 or Q3) the ratio needs be equal to or higher than 1.4. A similar ratio computed for the unstressed syllable (V2:C2) supposedly signals of Q3 if it is less than 1.4, while its values equalling or exceeding 1.4 indicate Q1 or Q2, without, however, discriminating between the two. (Eek, Meister 2003). Second, Eek has pointed out that the first syllable of a Q3 word should be distinguishable by its mean intensity of the first syllable being higher than that of the successive syllable. For Q2 the intensity of the two syllables is suggested as equal (Eek, Meister 1997).

3.2. Material and method

The material consisted of 485 words (including words of all three quantity degrees) read, in sentences, by 12 male and 13 female speakers of standard Estonian. Most of the samples analysed belongs to the Babel linguistic corpus in possession of the Institute of Cybernetics at Tallinn University of Technology, some additional samples had been read by two announcers from the Estonian Broadcasting Company. The recorded material was segmented and phonetically analysed by means of the PRAAT program (Boersma, Weenink 2008).

The research focus lies on vowel-centred Q1, Q2 and Q3 words where both the main stress syllable and the successive one have the structure CV(::)CV. In Q1 words the first-syllable vowel is short (e. g. *pole* [pole] 'is, are not'), whereas in Q2 and Q3 words the first-syllable vowel is long (e. g. *poole* [po:le] 'half GenSg' and *poole* [po::le] 'towards', respectively). Thus, the ratio of the stressed and unstressed syllables is found from the ratio of their vowel durations (V1:V2), while the ratio of adjacent phones in the first and second syllables can be written as V1:C1 and V2:C2, respectively. Most of the words are disyllabic, but longer words can also be found. A small number of the words begin either with a vowel or with a consonant cluster. The material also includes some principal and attributive components of compound words, and some foreign words pronounced like genuine ones; the total share of such words is less than a third of the whole bulk. The analysed material contains stressed as well as unstressed words from different positions (initial, middle, final) in the sentence or phrase. For each word its sound durations (ms) and the mean intensity of V1 and V2 (dB) were measured. The results were averaged, adding the standard deviation (SD).

3.3. Results

3.3.1. Duration ratios

The results of sound measurement and the ratios computed have been summarized in Table 1. (The total means and standard deviations have been calculated from the whole bulk of data, not from mean values.) In total the material contained 234 Q1 words, 150 Q2 words and 101 Q3 words.

V1:V2 is the classical ratio to be examined. The mean durations easily reveal that in Q1 words V1 is about twice as short as in Q2 and Q3 words, which is generally considered sufficient to perceive the short/long opposition between Q1 and the rest (the short Q1 vs. the long Q2 and overlong Q3).

Comparing our results with those received on laboratory speech earlier we find that our ratios are realistic, as far as quantity degrees go, albeit a little higher than expected for Q1 and Q2. In general, a typical duration ratio of Q1 should fall in the interval 0.6–0.7 (Lehiste 1960, Liiv 1961, Eek 1983, Eek, Meister 1997). Like the authors of the present study, G. Liiv as well as D. Krull analysed vowel-centred words. G. Liiv adds that a Q1 ratio can range from 0.50–1.00, while for Q2 the range is 1.00–2.00 (Liiv 1961). Most likely the reason for our slightly higher duration ratios for Q1 and Q2 words lies in that our research material of those two quantity degrees contains more foreign words and compound components.

For Q2 the ratios of V1:V2 vary more across different studies, ranging from 1.2–1.60; for Q3 words the range is 2.4–2.6 (Liiv 1961, Krull 1991, Eek ja Meister 1997). The vowel-centred structure has also been the research object for E. L. Asu and others, who study spontaneous speech. According to their results the average duration ratio is about 0.7 for Q1, 1.7 for Q2 and over 2.0 for Q3 (Asu a. o. 2009). Those numbers do not contradict ours either.

Next, let us consider the ratios of adjacent phones for the stressed vs. unstressed syllables. Our results for a stressed syllable confirm that V1 is short (as expected for Q1) if and only if V1:C1 is less than 1.4 (1.05 in Table 1). For Q2 and Q3 words the respective ratios are 1.99 and 2.59, respectively, which are both considerably higher than 1.4. In an unstressed syllable the ratio of the vowel V2 and its preceding consonant C2 is perhaps a little less unambiguous. Still, the theoretical ratio for Q1 and Q2 words being 1.4 or higher, our result for Q1 (1.8) should do well, and so does the ratio for Q2 as it can be rounded to 1.4 easily. Also, our V2-to-C2 ratio for Q3 words supports the theory as 1.17 is clearly lower than 1.4. Thus, our material indeed seems to corroborate Eek's theory.

Next, statistical methods will be used to find out which duration ratios, the traditional V1:V2 or the two-step system of V1:C1 → V2:C2 suggested by Eek, are

	C1	SD	V1	SD	V1:C1	SD	C2	SD	V2	SD	V1:V2	SD	V2:C2	SD
Q1	69	19	68	15	**1.05**	0.3	52	12	87	24	**0.82**	0.2	**1.74**	0.5
Q2	64	17	120	28	**1.99**	0.7	52	13	69	17	**1.80**	0.4	**1.37**	0.4
Q3	66	16	165	35	**2.59**	0.8	59	16	66	19	**2.59**	0.6	**1.17**	0.4

Table 1. Mean durations (ms), duration ratios and standard deviations (SD) of the first-syllable consonant (C1) and vowel (V1), the inter-vowel consonant (C2) and the vowel of the successive syllable (V2) in Q1, Q2, and Q3 words.

vital for classifying quantity degrees and, at the same time, more important for modelling speech temporal structure. Figures 5 and 6 present two CART-generated decision trees for quantity degree classification.

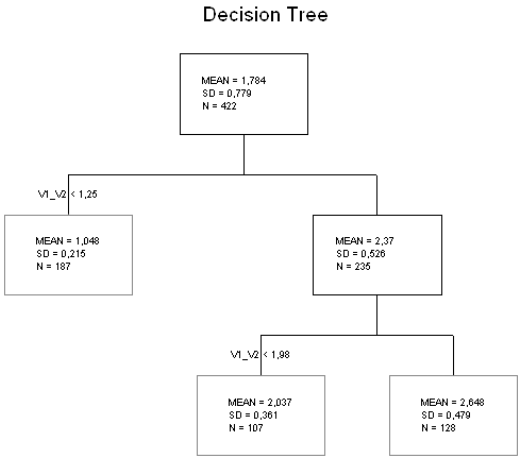


Figure 5. Decision tree based on the classical duration ratio of V1:V2.

From Figure 5 we can see that for Q1 the duration ratio V1:V2 is less than 1.25, while for Q2 the values of the ratio range from 1.25–1.98 and the criterion for recognizing Q3 is a V1-to-V2 ratio that exceeds 1.98. Figure 5 reveals that a decision tree based on two duration ratios (V1:C1 and V2:C2) actually manages to classify the quantity degrees by using only one of them (V1:C1) as the sufficient criterion. This points to the relative length (short or long) of the first syllable vowel V1 and thus, to the primary (durational) division of the quantity degrees into short (Q1) and long ones (Q2, Q3). Consequently, for our material the V2:C2 criterion has proved marginal after all. This brings back memories of M.Hint's theory of syllabic quantity degree, arguing that although phonetic quantity is manifested in the foot, its degree depends on certain parameters of the stressed syllable (Hint 2001).

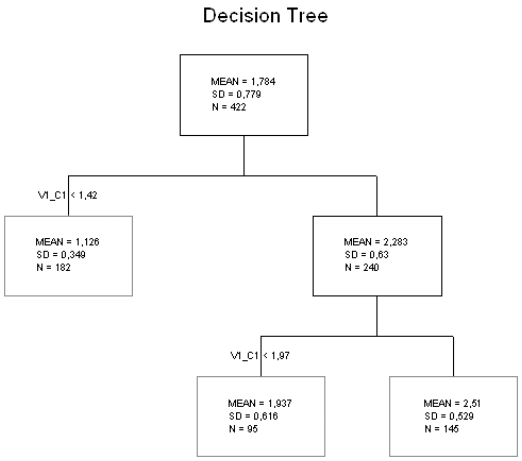


Figure 6. Decision tree based on a two-step system of the ratios V1:C1 and V2:C2.

For weighing the relevance of duration ratios in the model of phonetic quantity some simple equations of linear regression were generated. For classical duration ratios the linear model yielded quite a strong correlation between the input and output (correlation coefficient r=0.867). Consequently the model explains over 75 % of the data variation (coefficient of determination r^2=0.752). The alternative model generated from the other two duration ratios, however, yielded a correlation coefficient equal to 0.759, which means that it explains only 58 % of the variation in the data analysed.

According to the above results, the classical duration ratio (V1:V2) is the most relevant parameter to be considered in modelling the temporal structure of Estonian speech. However, although the above parameter guarantees quite a close correlation between the model input and output, similar tests should be run for some other physical factors, such as, for example, intensity and pitch, to find out their possible role in the formation of important phonological oppositions. In the present study the line will be drawn at intensity.

3.3.2. Intensity

Our analysis of intensity did not reveal anything remarkable. In all Q1 words the mean intensity of articulation is 73 dB, the only exception being the first vowel of Q1 words, which is pronounced at 74 dB. on average. The results for Q2 and Q3 offer more material for discussion and even confusion as according to our measurements the average intensity for Q2 words is 74 dB for V1 and 71dB for V2, while for Q3 words the respective readings are 74 and 72 dB. According to Eek and Meister, however, the average intensity of articulating the V1 and V2 in Q2 words is 73 dB throughout, whereas in Q3 words the respective intensities are 75 and 69 dB (Eek, Meister 1997). Our study has not detected so big a difference in any comparison; neither does it support the argument that V1 is articulated with most intensity in the initial syllable of Q3 words. True, there is a difference between the V1 and V2 as pronounced in Q2 and Q3 words (2–3 dB), respectively, observed in total as well as when comparing different word groups (stressed/unstressed, male/female), but the difference is not salient enough. Obviously intensity may help perceive the difference between Q2 and Q3 words, serving as a supportive feature if the duration ratio and the pitch curve, for some reason or other, fail to define unambiguously whether the word is a Q2 or Q3 one. Such cases are obviously very few. M. Parve, who analysed spontaneous dialect speech (Parve 2003) also reached the conclusion that although the intensity difference between Q2 and Q3 words can be distinguished by phonetic criteria, it is dubious or too vague for perception.

4. Conclusions

The conducted test of speech rate perception did not provide an unambiguous answer to all set questions. There are, indeed, certain differences observable between the speech rates preferred by the blind and the sighted, but the level of the difference depends not on the visual impairment but rather on the subjects' experience with using a computer and a screen reader. The ability to understand rapid speech appears after about three years of everyday practice of using of a computer and listening to synthetic speech. No so-called optimal speech rate can be established either for the blind or the sighted, as the preferred speech rate is very individual and depends on many circumstances.

The aim of the study was to find out whether Estonian quantity degrees could be distinguished by any other features but the traditional duration ratio of V1:V2. Our analysis of copious data proved that neither intensity nor adjacent sound ratio are as relevant as the ratio of the first and second syllable sounds. Possibly, the actual role of intensity in the quantity degree model could be approached better by pitch analysis. When modelling speech temporal structure one should keep in mind that standard Estonian is characterized by an alternation of words of three different quantity degrees, based on a natural alternation of stressed and unstressed syllables. The quantity degrees can be distinguished by a comparison of the duration ratios of those syllables, but obviously this is not all there is to it. The next object of research relevant in this respect should be the manifestation and role of pitch.

Acknowledgements

This work has been supported by the National Programme for Estonian Language Technology, grant ETF7998 and project SF0050023s09.

References

1. *Asakawa, C., Takagi, H., Ino, S., Ifikube, T.* 2003. Maximum Listening Speeds for the blind. Proceedings of the 20003 International Conference on Auditory Display. Boston, MA, USA, 276–279.
2. *Asu, E. L., P. Lippus, P. Teras, T. Tuisk.* 2009. The Realization of Estonian Quantity Characteristics in Spontaneous Speech. *Nordic Prosody — Proceedings of the Xth Conference*, Helsinki 2008. Editors Aaltonen, O., Aulanko, R., Vainio, M. Frankfurt: Peter Lang Verlag, 49–56.
3. *Boersma, P., D. Weenik.* Praat: doing Phonetics by computer (http://www.fon.hum.uva.nl/praat/)
4. *Eek, A.* 1983. Kvantiteet ja rõhk eesti keeles (I). Fonoloogiliste tõlgenduste kriitikat. — *Keel ja Kirjandus 9*, 481–489.
5. *Eek, A., E. Meister,* 1997. Simple Perception Experiments on Estonian Word Prosody: Foot Structure vs. Segmental Quantity. — *Estonian Prosody: Papers from a Symposium.* Editors Lehiste, I., Ross, J. Tallinn: Institute of Estonian Language, 77–99.
6. *Eek, A., Meister, E.* 2003. Foneetilisi katseid kvantiteedi alalt. — *Keel ja Kirjandus* 11–12, 815–837; 902–916.
7. *Hint, M.* 2001. Prosoodiaväitlustes läbimurdeta. — *Keel ja Kirjandus* 3–5, 164–172, 252–258, 324–335.
8. *Keller, Eric* 2007. Waves, beats and expectancy. — Proceedings of the 16th International Congress of Phonetic Sciences (eds. Jürgen Trouvain, William J. Barry). Saarbrücken, 6–10 August 2007. Saarbrücken, 355–360.
9. *Keller, Eric, Port, Robert* 2007. Speech timing: approaches to speech rhythm. — Proceedings of the 16th International Congress of Phonetic Sciences (eds. Jürgen Trouvain, William J. Barry). Saarbrücken, 6–10 August 2007. Saarbrücken, 327–329.
10. *Krull, D.* 1991. Stability in some Estonian duration relations. — Experiments in speech processes. PERILUS (Phonetic Experimental Research, Institute of Linguistics, University of Stockholm) No XIII, 57–60.
11. *Lehiste, I.* 1960. Segmental and Syllabic Quantity in Estonian. *American Studies in Uralic Linguistics Vol 1*. Bloomington, Indiana University, 21–28.
12. *Lehiste, I.* 1997. Search for phonetic correlates in Estonian Prosody. — Estonian Prosody: Papers from a Symposium, Proceedings of the International Symposium on Estonian Prosody, Lehiste, I.; Ross, J. (eds.). Tallinn, Estonia, October 29–30, 1996. Institute of the Estonian Language and Authors, Tallinn, 11–35.
13. *Liiv, G.* 1961. Eesti keele kolme vältusastme kestus ja meloodiatüübid. — *Keel ja Kirjandus* 7–8, 412–424, 480–490.
14. *Mihkla, M.; Meister, E.* 2002. Eesti keele tekst-kõnesüntees. Keel ja Kirjandus, 45(2); 88–97 ja 45(3); 173–182.
15. *Moos, A., Trouvain, J.* 2007. Comprehension of Ultra-Fast Speech — Blind vs „Normally Hearing" Persons., 677–680.
16. *Moos, A., Hertrch, I., Dietrich, S., Trouvain, J., Ackermann, H.* 2008. Perception of Ultra-Fast Speech by a Blind Listener — Dos He Use His Visual System? Proceedings of the 8th Seminar on Speech Production, ISSP 2008, 297–300.
17. *Parve, M. 2003.* Välted lõunaeesti murretes. Doktoritöö. Dissertationes philologiae estonicae universitatis tartuensis 12. Tartu: Tartu Ülikooli Kirjastus.
18. *Ross, J. Lehiste, I.* 2001. The temporal Structure of Estonian Runic Songs. *Phonology and Phonetics 1.* Editor A. Lahiri. Berlin; New York: Mouton de Gruyter.

Управление лексиконом в онтологической семантике

Lexicon management in ontological semantics

Petrenko M. (mpetrenk@gmail.com)

Московский Гуманитарный Институт им. Е. Р. Дашковой. Москва, Россия;
RiverGlass, Inc. Champaign, IL, USA.

В работе рассматриваются способы управления лексиконом — одним из базовых (наряду с онтологией) статических ресурсов в Онтологической Семантике. Описывается общая стратегия освоения лексикона (lexicon acquisition), описывается ряд техник освоения, и, на примере освоения английских глаголов класса с инструментально-субъектной альтернацией, описывается, как решается проблема освоения проблемных лексических единиц.

1. Paper goals

The paper describes how the lexicon — a static knowledge resource — is managed by a human acquirer. The study draws on the methodology, theory and strategy of lexical acquisition outlined in [3] and takes into account the ongoing implementation experience in various applications, as well as recent revisions/improvements. After a brief outline of the lexicon, the general strategy of lexical acquisition will be introduced, and techniques of acquisition described. An example will then illustrate how complex cases are handled through lexical acquisition within the framework of the Ontological Semantic Technology (OST).

2. Ontological Semantics: static knowledge resources

The architecture of Ontological Semantics, also known as Direct Meaning Access, comprises a set of static and dynamic resources. The Ontological Semantics school of thought subscribes to the semantic prerequisite in NLP and is premised on the idea that the full (i. e. human-like) efficiency in natural language processing is only attainable through a solid knowledge resource base, which would (1) model the world as a complex and highly structured conceptual hierarchy and (2) furnish lexical, morphological, and syntactic knowledge essential for parsing a natural language input meaningfully (for a more detailed discussion and support of the need to "do semantics semantically", see [1], [7], and [8]).

2.1. The ontological knowledge resource

A detailed and in-depth description of the ontology is offered in [Taylor et al in this volume], so this subsection will contain only a very brief outline. The concepts of the ontology enter into a large number of relations: the hypero-hyponymic (i. e. class-subclass) relation branches the root concept ALL into EVENT, OBJECT, and PROPERTY. Breaking further into numerous subclasses, EVENT's take a large number of properties (including, but not limited to, case-roles) filled by OBJECT's. Both OBJECT's and EVENT's are, in turn, defined through a broad spectrum of circa one hundred ATTRIBUTE's and RELATION's within and across their branches. As illustrated by the example below, the concept BOX is defined through (i. e. is in the DOMAIN of) the properties MADE-OF, CONTAINS, and SHAPE. The concepts CERAMIC, METAL, PAPER, PLASTIC, WOOD function as fillers (i. e. are in the RANGE) of the property MADE-OF.

(box
  (definition (value("a rectangular container")))
  (is-a (hier(container)))
  (**made-of** (sem(**ceramic metal paper plastic wood**)))
  (shape (value(rectangular square)))
)

Which concepts can fill which properties is regulated by the restrictions on the properties' DOMAIN and RANGE. In other words, the concept BOX can take the property MADE-OF because its ancestor, PHYSICAL-OBJECT, fills the DOMAIN of MADE-OF. The ontological

entry for concept PHYSICAL-OBJECT with hard-coded properties is provided below:

```
(physical-object
   (definition  (value("objects that physically exist")))
   (is-a        (hier(object)))
(subclasses     (hier(surface-feature landscape-object
                animate animate-part animal-artifact
                material artifact celestial-object)))
)
```

The concepts CERAMIC, METAL, PAPER, PLASTIC, and WOOD can fill the RANGE of MADE-OF because their ancestor, MATERIAL, fills the RANGE of MADE-OF, as illustrated by the example of the concept MADE-OF below.

```
(made-of
   (definition  (value("the relation between a thing and
                things made out of it")))
   (is-a        (hier(physical-object-relation)))
   (inverse     (value(material-of)))
   (domain      (sem(physical-object)))
   (range       (sem(material)))
)
```

2.2. The lexical knowledge resource

The lexicon is a language-specific repository of word senses coupled with their morphological and syntactic information (for a detailed description of the template of a lexicon entry see [3]). As a static knowledge resource and a resource used directly by the OST text parser, the Lexicon fulfils two important functions.

In relation to the ontology, the main function of the Lexicon is to map the language-independent ontological knowledge to syntactic and semantic features of a specific language, including semantic idiosyncrasies. The mapping procedure can be either direct anchoring, if the ontology has a concept that exhaustively captures the meaning of a lexical sense (illustrated below by the entry "car-n1") or indicating the semantically nearest (typically a class) concept and specifying its meaning through properties and their restricted fillers (illustrated below by the entry "tourist-n1").

```
(car-n1
      (cat(n))
      (synonyms "")
      (anno(def "an automobile")(comments "")
            (ex "he drives a car"))
      (syn-struc((root($var0))(cat(n))))
      (sem-struc(car)))
(tourist-n1
      (cat(n))
      (synonyms "")
      (anno(def "a person who travels")(comments "")
            (ex "the tourists stayed at the hotel"))
      (syn-struc((root($var0))(cat(n))))
      (sem-struc(human(agent-of(sem(travel))))))
```

In relation to the OST text parser, the main function of the Lexicon is to provide the OST text parser with essential data about the word sense, its syntactic position and semantic information in the sentence so that the machine could (1) retrieve the proper ontological information about the sense, including concepts, their properties and property fillers, and (2) by computing property fillers, accommodate the given sense a text-meaning representation (TMR) of the natural language input.

The TMR is the ultimate product of the OST text parser. It comprehensibly translates a natural language input into a configuration of semantically related concepts, as illustrated by the below example of a TMR of the sentence, "the tourist broke the box", where the concepts HUMAN(agent-of(sem(travel))) and BOX fill the case roles of agent and theme of the clause-forming event DAMAGE. A more in-depth explanation of how TMR are computed within OST is offered in [4], Sections 6 and 8 (see also [3] and [5]).

(110) *The tourist broke the box.*

```
TMR 1:Weight: 4.2 Event: break-v1,
      damage1    agent(value (tourist-n1,
                 human1(agent-of(sem(travel)))))
                 theme(value (box-n1, box1 ))
```

3. Lexical acquisition

The practice of lexical acquisition involves the machine-readable description of every lexical sense in a domain-specific corpus, following the principle of complete coverage stated in [3] (Section 9.3). A well-acquired lexical entry (1) is anchored in an appropriate concept (which is evidenced from the concept's location in the ontological hierarchy, its ancestors, siblings, descendants, and its ontological and prosaic definition), (2) matches syntactic and semantic structures through properly co-indexed variables, and (3) reflects all possible syntactic positions the word may take in the sentence (for more details on steps of lexical acquisition see [3], Section 9.3.4). Procedurally, two lexical acquisition strategies are outlined in [3] (Sections 9.3.2 and 9.3.3). The first strategy, acquisition by rapid propagation, involves covering a large class of semantically and syntactically similar entries by applying, with slight modifications, one "master" lexical template. The degree of modification varies depending on the class size and the homogeneity of its members: while the acquisition of scalar adjectives would mostly require only changing the head concept ATTRIBUTE and its numeric value in the sem-struc, the acquisition of regular nouns like "car-n1", deverbal nouns like "investigation-n1", and deadjectival nouns like "beauty-n1" requires a greater

degree of syntactic and semantic variation. The second strategy is acquisition based on lexical rules of converting grammatical cognates like verbs (e. g. “enjoy-v1”) and their adjectival derivatives (e. g. “enjoyable-adj1”) due to their semantic similarity.

In order to facilitate various aspects of an application, several techniques of lexical acquisition may be defined.

1) Ontology-driven lexical acquisition involves “sliding” down the ontological hierarchy and making sure all concepts of the OBJECT and EVENT branches have minimal representation in the lexicon. While this technique is time-efficient and quickly produces a workable lexicon, its obvious downside is the limited size of the lexical entries, each of which will most likely have only one sense. This technique could be employed at the early stage of ontological acquisition, when the ontology is not yet complete, so that “lexicalizing” the concepts early on would make both resources available for parser-based testing, which is most beneficial in the overall ontology assessment and often points to necessary adjustments.
2) Parser-driven lexical acquisition involves running the OST text parser on a large number of domain-unrestricted corpora. Analyzing the resulting TMR’s allows establishing whether an additional lexical entry needs to be introduced or if it is the existing entry that has not parsed, in which case an adjustment is required. A properly conducted TMR analysis (informally known in the OST community as “blame-assignment”) also helps identifying whether the processing issues are rooted in the ontology, the onomasticon, or dynamic parsing modules.
3) Domain-driven lexical acquisition involves running the OST text parser on a domain-specific corpus. The corpus size and the depth of parsing are largely determined by the application purposes. The application also establishes the focus (e. g. grammatical classes) and the grain size (number of senses per entry) of lexical acquisition. To further fine-tune the acquired corpus to a specific domain, the priming functions can be introduced that prime (a) a lexical sense within the entry based on its general regularity in the language, and (b) a domain-specific lexical sense. This acquisition technique works best when aided by a pre-processing module comprising a tagger, a stemmer and a look-up function, which compares the corpora to the lexicon and identifies missing lexical entries with further part-of-speech sorting.

A usual build out of an application typically involves the interaction of the three lexical acquisition techniques described above. While technique (1) is largely restricted to early development phases, (3) is heavily guided by immediate objectives, (2) constitutes the backbone of lexical acquisition. When applied on a more limited scale and a case-by-case basis, this technique can also be used to test the functionality of every newly acquired or adjusted lexical entry. This is done by running a sample sentence (drawn from a corpus or emulated) with the new entry through the OST text parser, and analyzing the resulting TMR. The example below illustrates a typical lexical acquisition cycle supported by the TMR analysis.

Let us assume that a corpus related to the domain of crimes contains a sentence: “The police arrested the mole for stealing data from federal servers”. The running of the OST text parser returns no TMR for this sentence. For the sake of clarity, let us initially focus on the first clause, “the police arrested the mole”. The analysis starts by looking up the lexicon entries for “arrest-v” and “mole-n”:

```
(arrest-v1
    (cat(v))
    (anno(def "to seize a person by legal authority or warrant")(comments "")(ex "the police arrested the arsonist"))
        (synonyms "")
    (syn-struc(
        (subject((root($var1))(cat(np))))(root($var0))(cat(v))
        (directobject((root($var2))(cat(np))))))
    (sem-struc(arrest
        (agent(value(^$var1)))
        (beneficiary(value(^$var2(should-be-a(sem(human))))))))))
```

The lexicon has the following entries for the word “mole”:

```
(mole
    (mole-n1
        (cat(n))(synonyms "")
        (anno(def "a insectivorous mammal living underground")(ex "he noticed a mole in the ground"))
        (syn-struc((root($var0))(cat(n))))
        (sem-struc(rodentia(agent-of(sem(life-event(location(sem(soil))))))))))
    (mole-n2
        (cat(n))(synonyms "")
        (anno(def "a spot on the skin")(comments "")(ex "the man injured the mole"))
        (syn-struc((root($var0))(cat(n))))
        (sem-struc(skin(relative-size(less-equal(0.3)))(color(value(black brown)))))))
```

(111) *The police arrested the mole.*

```
TMR 1: Weight: 2.12 Event: arrest-v1,
                           arrest1
                           agent(value (police-n2, police-officer1 ))
                                 beneficiary(value (mole-n3, human1(agent-of (sem(spying)))))
```

(112) *The police arrested the mole for stealing data from federal servers*

```
TMR 1: Weight: 6.31 Event: arrest-v1,
                            arrest1
                            agent(value (police-n2, police-officer1 ))
                                  beneficiary(value (mole-n3, human1(agent-of(sem(spying)))))
                                  precondition(value (steal-v2, larceny
                                                theme(value (data-n1, information (origin(value (server-
                                                      n1, computer(connected-to(sem(network))))))))))
                                                      owned-by(value (government, federal-adj1))
```

None of the head concepts RODENTIA or SKIN in the sem-strucs of "mole-n1" or "mole-n2" can fill the beneficiary case role of ARREST, which is constrained to HUMAN according to the entry "arrest-v1". A lexicon acquirer would then conclude that a lexicon sense of "mole-n3" is needed which would (1) comprehensibly describe the meaning of the word "mole" as "a double agent, spy" and (2) have a head concept in its sem-struc that could fill the beneficiary case role of "arrest-v1"[1]. An ontological lookup will have no direct concept for SPY[2], so the nearest class concept will be listed in the sem-struc with the constraining property (agent-of(sem(spying))). To check whether this description is warranted by the ontology, the concept SPYING will be checked for its AGENT fillers. No restrictions are listed for the AGENT of the concept SPYING, which means that the machine will find the AGENT filler from the RANGE of the property AGENT in the ontology, and this filler is ANIMATE. Since HUMAN is a descendant of ANIMATE in the ontology, the sem-struc (human(agent-of(sem(spying)))) is supported by the ontology. The resulting sense will have the form:

```
(mole-n3
    (cat(n))
    (synonyms "")(anno(def "a double agent") (ex "the
        mole was arrested"))
    (syn-struc((root($var0))(cat(n))))
    (sem-struc(human(agent-of(sem(spying)))))
)
```

Re-running the clause with the OST text parser, would return the TMR [example (2)]:

In case the issues persist (e. g. no TMR is returned, the TMR is not correct, etc.) a more thorough insight into the output of every module would be needed, starting from the pre-processing steps of part-of-speech tagging and stemming.

The processing[3] of the second clause, "for stealing data from federal servers", will involve the clause-merging module of the OST parser. The module will do a lexical lookup of the preposition "for" and locate a sense "for-prep4", which is anchored in the property PRECONDITION (a child of EVENT-RELATION), whose DOMAIN and RANGE, in turn, have EVENT's as fillers. The two clauses will thus be merged into (arrest(precondition(sem(steal)))). The preposition processing module will be activated to process the noun phrase "data from the servers": the entry "from-prep1" will map on the property ORIGIN (a child of OBJECT-RELATION), whose DOMAIN and RANGE will be filled with INFORMATION and COMPUTER (identified through the lexical entries "data-n1" and "server-n1", respectively). The adjective processing module will be called to parse the adjectival phrase "federal servers": the property OWNED-BY (a child of SOCIAL-OBJECT-RELATION) will be located through the lexical entry "federal-adj1", its DOMAIN will be found to match the concept COMPUTER of the modified noun "server-n1", and the range filler GOVERNMENT (a child of ORGANIZATION) from the semantic structure of "federal-adj1" will be copied into the TMR for the concept COMPUTER. The resulting TMR of the whole sentence will have the form [example (2)]:

---

[1] Condition (1) clearly prevails over (2) since efficient processing is the ultimate goal of the system, and if the meaning is described accurately and the ontology cannot accommodate it, ontological adjustment is in order.

[2] The issue of a balanced trade-off in distributing knowledge between the ontology and the lexicon has been discussed in [3] (see also [7]). Whenever a lexical entry has no direct anchoring concept in the ontology, the decision whether a new concept should be added is guided by (1) the considerations of the parsimony of the ontology, which is a language-independent construct; (2) the purposes of a specific application, which defines the grain size of ontological and lexical acquisition.

[3] The author is grateful to the anonymous reviewers for emphasizing the need to illustrate/elaborate on the functionality of the Ontological Semantic Technology based on real-life data. The example contains clause embedment, prepositional phrase and an adjectival modifier, and its parsing would require the deployment and integration of several task-specific modules based on rich ontological and lexical knowledge resources.

4. Handling problematic cases in lexical acquisition in OST

The section below will describe how problematic cases can be acquired with the lexical acquisition inventory. More specifically, the lexical acquisition of verbs with Instrument-Subject alternation will be discussed.

A class of verbs exists where the event can be carried out through an agent or instrument [3].

(113) [illegible]

(114) [illegible]

(115) [illegible]

(116) [illegible]

While different solutions were proposed to further stratify the instrument case role into intermediary or facilitating types or relax the notion of subject to include instrumental subjects [2, p. 80], within the framework of OST, the issue translates into the question of how to interchangeably accommodate two distinct case roles of AGENT, INSTRUMENT, and the relation PRECONDITION in one syntactic position indexed by a variable in the syn-struc of an entry like "break-v1".

In the ontology, the case role of AGENT has its RANGE restricted to ANIMATE, which rules out HAMMER (a descendant of ARTIFACT), ASTEROID (a descendant of CELESTIAL-OBJECT), and HURRICANE (a descendant of PHYSICAL-EVENT). On the other hand, the INSTRUMENT case role does not have an animate object in its RANGE, and the PRECONDITION is an EVENT-RELATION, which excludes any OBJECT by definition. Acquiring three separate lexical senses for AGENT, INSTRUMENT and PRECONDITION is not entirely justified: all three entries would have shared the concept DAMAGE and would have been identical syntactically.

A reasonable solution would be to expand the sem-struc of the entries like "break-v1" to include additional case roles that would map on one syntactic variable, which would result in the following entry:

```
(break-v1
    (cat(v))(anno(def "to cause to break")
    (ex "He broke the window. The hammer broke the window. The hurricane broke the window.")(comments ""))
        (synonyms "")
    (syn-struc ((subject((root($var1))(cat(np))))(root($var0))(cat(v))
    (directobject((root($var2))(cat(np))))))
    (sem-struc (damage(agent(value(^$var1)))
    (instrument(value(^$var1(should-be-a(sem(artifact animate-part material celestial-object)))
    (precondition(value(^$var1))))))
    (theme(value(^$var2(should-be-a(sem(artifact)))))))))
```

Such an entry conforms to the ontological restrictions, because the concepts ARTIFACT, ANIMATE-PART, MATERIAL, CELESTIAL-OBJECT constraining the instrument case role of "break-v1" are within the RANGE of the INSTRUMENT in the ontology, and the concept ARTIFACT constraining the theme case role of "break-v1" is within the RANGE of the property THEME in the ontology. The unconstrained case roles of agent and precondition in "break-v1" will be restricted by the RANGE of the property AGENT (which is ANIMATE) and the RANGE of the property PRECONDITION (which is EVENT).

When reading an entry above during the processing of examples (4–7), the OST text parser would selectively fill (and display in a TMR) the agent case role with HUMAN in (4), the instrument case role with HAMMER in (5), the instrument case role with ASTEROID in (6), and the precondition relation with HURRICANE in (7). The following TMR's will thus be derived:

(117) *The man broke the window*

```
TMR 1: Weight: 4.2200003 Event: break-v1,
                                 damage1
                                        agent(man-n1, human1(gender(value(male))))
                                        theme(value (window-n1, window1 ))
```

(118) *The hammer broke the window*

```
TMR 1: Weight: 4.2 Event: break-v1,
                          damage1
                                 instrument(value (hammer-n1, hammer1 ))
                                 theme(value (window-n1, window1 ))
```

(119) *The asteroid broke the window*

TMR 1: Weight: 4.16 Event: break-v1,
damage1
instrument(value (asteroid-n1 asteroid1))
theme(value (window-n1, window1))

(120) *The hurricane broke the window*

TMR 1: Weight: 4.18 Event: break-v1,
damage1
precondition(value(hurricane-n1 hurricane1))
theme(value (window-n1, window1))

The lexicon thus offers a very versatile toolbox for acquiring complicated word classes comprehensively. The rich ontology allows for a correct representation of semantic multiplicity as separate lexical senses. An exhaustive descriptive vocabulary of syntactic properties helps to accommodate syntactic variation in a single lexical entry, which keeps the lexicon meaningfully parsimonious. At the management level, the three acquisition techniques described above make it possible to calibrate the scope and grain size of acquisition to a specific task and based on a specific application. Lexicon acquisition informed by the OST text parser provides a most balanced and illuminating approach to quality control and improvement.

References

1. *Hempelmann C. F., Raskin V.* Semantic Search: Content Vs. formalism // Rome: Proceedings of Langtech 2008. http://www.langtech.it/en/technical_program/technical_program.html (full paper).
2. *Levin B.* Verb Classes and Alternations: A Preliminary Investigation // London and Chicago. The University of Chicago Press, 1993.
3. *Nirenburg S., Raskin V.* Ontological Semantics // Cambridge, MA: MIT Press, 2004. A prepublication draft, chapter by chapter, can be found on www.ontologicalsemantics.com
4. *Petrenko M.* 2009. Ontological semantics and abduction: parsing ellipsis.// Dialog 2009.
5. *Petrenko M., Raskin V.* Modeling Abduction within Ontological Semantics // Proceedings of Midwestern Computational Linguistics Colloquium 5. East Lansing: Michigan State University, May 2008.
6. *Raskin V.* The how's and why's of ontological semantics. In: I. M. Kobozeva, A. S. Narinyani, and V. P. Selegei (eds.) // Zvenigorod: 2005. Computer Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2005".
7. *Raskin V., Hempelmann C. F. and Taylor J. M.* Guessing vs. Knowing: The Two Approaches to Semantics in Natural Language Processing. // In this volume, 2010
8. *Raskin V., Hempelman C. F., Taylor J. M., Petrenko M. S., Trienzenberg K. E., Buck B.* The Why's, How's, and What-of's of Natural Language Ontology // Meaning Computation 1 (forthcoming), 2010.

Догадка или знание: два подхода к семантике при обработке естественного языка

Guessing Vs. Knowing: The Two Approaches to Semantics in Natural Language Processing

Raskin Victor (vraskin@purdue.edu),
Hempelmann Christian F. (khempelmann@riverglassinc.com),
Taylor Julia M. (jtaylor1@purdue.edu)

Purdue University & RiverGlass Inc, USA

В статье рассматривается принципиальное различие между двумя подходами к семантике при обработке естественного языка. Первый пытается установить статистически, по совместному употреблению слов как цепочек без значения, о чем говорится в тексте. Авторы предпочитают противоположный подход, построенный на прямом и полном представлении значения текста.

1. Preface: The Brief History of the Dichotomy

This paper discusses two approaches to semantics in natural language processing (NLP), the prevalent statistical/machine learning approach (SML) and the persistent meaning-based minority approach, the "real" computational semantics, based on direct and comprehensive meaning access (DMA). Roughly the same division is captured by contrasting SML to the rule-based approach. Yet another way to expose the dichotomy is to separate the meaning and usage of NLP and computational linguistics—contrary to the common practice of using the two terms synonymously. And, finally for now, the difference is often placed in the attitude towards the "knowledge acquisition bottleneck" (Navigli 2009—see also Section 4 below), with SML recognizing it and dismissing knowledge acquisition as subjective, non-scalable, and non-feasible, and DMA making it the cornerstone of its endeavor.

This dichotomy is not new: there is a great deal of similarity between its current guise and the naïve "95 % vs. 100 %" debate in early machine translation (MT: Akhmanova *et al*. 1961), with the "100 %" side insisting on knowing and understanding everything before computing it, and the "95 %" side interested in trying all kinds of formal methods, from an extremely limited repertoire then available, and finding out what they can tell us. Bar-Hillel (1954) introduced the notion of the semantic bottleneck on the way to a fully automatic high-quality MT very early on, and in the next decades, whether informed by his opinion or not, most people in NLP tried non-semantic methods while a small minority attempted to remove the bottleneck. The former, many of them linguists, but not semanticists, attempted to perfect the syntactic parsers, thinking that this would somehow make semantics unnecessary.

The US government funding for MT disappeared with the publication of the "Black Book" (Languages and Machines 1966) and resumed slowly in the mid 1980s with a new, knowledge-based approach. Other applications, such as information retrieval, information extraction, text mining, summarization, search, and question answering were emerging. By 1990 or so, the syntacticians were replaced, in the same non-, a-, and outright anti-semantic camp, by people, mostly non-linguists, who were interested in applying statistical and machine-learning methods instead of syntax, first admitting openly and even bragging about fully sharing the non-semantic orientation of their predecessors and then, since roughly around 1997, when most government funding started stipulating the necessity of semantics, claiming that theirs was the only way to the meaning of texts. On the other side, a small group of computational semanticists was making a case for the feasibility of acquiring semantic resources, such as the ontology and lexicons.

Another dimension bearing on the same dichotomy is the relation between NLP and artificial intelligence (AI). During the peak of excitement about AI, in the 1980s, NLP proudly referred to itself as the NL AI. Against the background of the Chomskians' (typically, much more stridently than Chomsky's himself) claims that the transformational generative grammar, which was mostly syntax, was not just a model of what was internalized in the mental mechanisms underlying lan-

guage but actually the content of those mechanisms (see a critique in Raskin 1979), the "just a model" position allied itself with the "weak AI" principle; the "content" position with the strong AI thesis. Very few NLP principals adopted the latter, which claimed that it would develop machines which would think exactly like humans—or, conversely, that the principles, on which those machines will be built, will be identical with the principles on which the human mind operates. But the weak AI thesis was very appealing and more realistic—and compatible with the black box basis of cybernetics at its inception.

NLP systems may be seen in two different lights depending on whether their developers are interested in their results, no matter what methods are applied, or in the methods emulating human thinking. The SML methods can be easily seen as not motivated by the weak AI thesis while the DMA methods appear clearly informed by it. Simplistically and not fully accurately, people process language by knowing the meanings of the words and phrases and how they are combined together in sentences—they do not process large massives of statistical data, certainly not rapidly enough for that to be useful for instantaneous understanding of text.

Finally, an interesting historical parallel invites itself into this discussion. Late in the 19th century, when Frege (1884) introduced the preoccupation with the foundations and justification of well-established disciplines that has led to the emergence of the philosophy of science and the philosophies of specific disciplines, Russell wanted the exact definitions of various terms. Having failed to find them from the linguists because linguistic semantics, barely 40 years old, had not yet graduated to meaning definitions, he wrote it off and developed the parallel discipline of the philosophy of language to deliver what he needed (it has not). In the 1990s, when meaning became a *sine qua non* of NLP, the computer scientists and statisticians could not find any semantic help in the field and proceeded to develop the non-linguistic methods to do something apparently meaning-related.

2. The Plane of Expression vs. the Plane of Content

The distinction between what is perceivable by the senses and what it can symbolize, always known to humans and possibly to animals, had received a conceptual support in philosophy (Peirce 1991), from which the new discipline of semiotics emerged, before it reached linguistics in the writings of Saussure (1916): his *signifiant* was the material side of the sign that people can hear or see, while his *signifié* is the meaning of the sign. His maverick European successor, Louis Hjelmslev (1953), captured the same distinction in his plane of expression for the *signifiant* and plane of content for the *signifié* and bifurcated each plane into form and substance, thus imposing structure on both planes and introducing the notion of commutation between the two, so that the changes on the one plane are reflected on the other. He then proceeded to remove the two substance disciplines, phonetics (substance of expression) and semantics (substance of content) from his linguistics, leaving only phonology (form of expression) and grammar (form of content) in. Deviating from his own decree, however, Hjelmslev later (1958) made one of the very few structuralist contributions to semantics by trying to impose structure on content by comparing the dissection of the same narrow semantic fields (*tree/wood/forest*; *brother/sister/sibling*) in multiple languages (see also Trier 1931, Weisgerber 1950).

It is interesting to throw the two-plane perspective and their commutation on NLP. The syntacticians of the 1960–80s, quite independently of Hjelmslev, who did not leave a school and was—and is—largely forgotten, also attempted to get to the content through its form. These efforts, reinforced first by Chomsky's (1965) venture into a tiny area of semantics, immediately adjacent to syntax (and annexed by him into syntax) and then by a massive infusion of first-order predicate logic, rediscovered by linguists yet again, probably in part thanks to the elegant McCawley (1993), led to the separation of formal semantics from lexical semantics and to the exclusive focus in the thus redefined discipline of semantics on the former (the attempts to save lexical semantics from the charge of being just substance by discovering the grinding rule—*cow/beef, sheep/mutton*—were entertaining but short-lived: see Nirenburg and Raskin 2004: 117 and references there). At its peak, however, formal semantics had to admit that syntactic distinctions and semantic distinctions did not coincide (see Raskin 1994 and references there) but continued to operate, often with admirable virtuosity, on this counterfactual basis and to focus on the most grammatical aspects of meaning, such as quantifiers and other direct reflections in NL of what was clearly defined in logic.

SML, however, operates on the unarticulated assumption that substance is not accessible directly at all, form or content. Instead, it believes that the regularities of co-occurrence, masterfully augmented on multiple parameters, can classify texts without actually understanding its meaning by the computer. Whether they know it or not (and some do), they operate in the Wittgensteinian "language is usage" tradition reflected in semantics by Firth's (1957, cf. Raskin 1971) meaning by collocation: an important part of the meaning of *dark* is its collocation with *night*, and vice versa.

Contrary to that, DMA believes that direct and comprehensive (non-selective) approach is essential for the ultimate success of NLP's growing list of applications. In plain linguistic terms, it means the delivery of the meanings of words and phrasals to the computer. Their opponents may even agree to this position—they simply claim that it is impossible to accomplish (see Sec-

tion 5 below). In the next two sections, we will sketch out the alternative positions of the two camps. We will then discuss the evaluation attack by the one-planer SML on the two-planer DMA, and the need for the counter-offensive.

3. SML: A Loving View

It is not that the "non-semantic" approach is not interested in the semantics of text: one has to be if one is to compete these days for NLP funding. It is just that they, *a priori*, consider unimplementable the universal program of linguistic semantic theory, which, since and after the much maligned but seminal Katz and Fodor (1963), has included a lexicon and the compositional rules combining lexical into sentential meanings. Instead, they use a combination of statistical methods (Manning and Schuetze 2000) with machine learning (Mitchell 1997).

Because, as mentioned above, ambiguity has been seen as the main cause for Bar Hillel's semantic bottleneck, it is useful to see how this approach handles the problem of word sense disambiguation (WSD: see Kilgariff 1997, 1998, 2006; Ide and Veronis 1998; Ide and Wilks 2006; cf. a very useful and candid survey in Navigli 2009 and references there). We will describe here a common, most typical, generic case rather than any particular implementation, so various improvements may have been already added in later manifestations, without affecting this discussion.

A large corpus of text is divided into a smaller training part and a test part. Selective ambiguous (polysemous or homonymous) words are marked throughout the text, usually no more than one per sentence, and their different sense, commonly no more than two per word, are indicated. Human subjects mark the sense that they prefer contextually. Then the statistical/machine learning system attempts to guess the correct senses of the similarly marked words in the test corpus on the basis of statistical properties it observes in the contexts of the selections, mostly the co-occurrence of certain words significantly more frequent than its random probability. This principle has not changed basically but has been much refined since such early pioneering works as Shaykevich (1963).

A typical application for WSD has been a search for documents from a large control corpus pertaining to a small set of keywords, and the results are evaluated in terms of recall and precision. Recall is the number of hits retrieved divided by all hits in the corpus that should have been retrieved, and precision is the number of hits retrieved divided by the number of all instances retrieved, including false positives. The evaluation metrics are an important part of the approach (see also Section 5 below), and they have been perfected in multiple SemEval/SenseEval (see, for instance, Aguirre *et al*. 2007) competitions that are part US Government-stipulated for all of its grantees and part voluntarily participation by the proponents. The improvement on these scores gives a sense of individual and industry progress; it translates into rankings, awards, prizes, etc. The practitioners of the approach see these evaluations as the NLP standard and attempt to assess other approaches in these terms.

The approach avoids the challenging problems of understanding text beyond judging it pertinent to a keyword set and thus ignores all the problems of understanding how natural language works and how the human processes information. Unattested input, plasticity of meaning, salience, inference, reasoning in NL do not present any problem for it, and the practitioners are proud of it. It is widely believed that even WSD is "an AI-complete problem [Mallery 1988], that is, by analogy to NP-completeness in complexity theory, a problem whose difficulty is equivalent to solving central problems of *artificial intelligence* (AI), for example, the Turing Test[Turing 1950]" (Navigli 2009). He continues to say that "[u]nfortunately, the manual creation of knowledge resources is an expensive and timeconsuming effort [Ng 1997], which must be repeated every time the disambiguation scenario changes (e.g., in the presence of new domains, different languages, and even sense inventories). This is a fundamental problem which pervades the field of WSD and is called the *knowledge acquisition bottleneck* [Gale et al. 1992b]."

In the next section, presenting DMA, we will respectfully question these statements and attribute them to the lack of interest, ideological and disciplinary preferences, and in many cases, insufficient understanding of linguistic semantics and/or experience with descriptive semantics. We will have to agree, regretfully, with Navigli's rather damning (to us) self-assessment that "[t]he hardness of WSD [in SML practice—added by us] is also attested by the lack of applications to real-world tasks."

4. DMA: A Critical Self-Scrutiny

As described in Sections 1–2, the opposite approach, DMA, to which we refer as semantic and sometimes, for emphasis, "semantic semantic" or "real semantic" (see Hempelmann and Raskin 2008), differs in that it: (a) sets itself and the field the much more ambitious goal of direct and comprehensive meaning access to text and (b) proceeds to that goal, in full compliance with the program of linguistic semantic theory, by acquiring the same knowledge resources as it sees the humans as having and using in understanding NL. It also fully subscribes to the weak AI thesis, thus pursuing an interest in modeling/emulating the way humans are hypothesized to process meaning. In other words, besides the interest in getting optimal results in NLP applications,

this approach maintains that the only way to achieve this is by programming the computer to emulate human understanding. Capitalizing on the training and experience in descriptive linguistic semantics, it is free of the "fear of semantics" (Raskin 1988) and directly attacks what it does not actually perceive as the "knowledge acquisition bottleneck." It disagrees with the AI-complexity assessment of its task and believes that, while actually fraught with more complexities than SML recognizes, the goal is implementable, and it will suffice to do it only once, with the methodology of domain extension in place and robustness with regard to unattested input built in.

We will illustrate this on the example of a specific approach we have been developing, improving, and implementing for a couple of decades. Ontological Semantics has been extensively reported in a number of publications since its inception in the late 1980s, most comprehensively in Nirenburg and Raskin (2004) and recently at this forum as well (Raskin 2006, Petrenko 2009). The current incarnation of Ontological Semantics, which we call the Ontological Semantics Technology (OST) is characterized by the implementation of many but not all features outlined in Nirenburg and Raskin (2004), resulting in a considerable revision of these features and sometimes radical departures from the earlier views. A proprietary implementation of the system has reached a functional implementation stage, well beyond the earlier academic proof-of-concept demonstrations, such as the MikroKosmos MT system (see the references in Nirenburg and Raskin 2004: 29).

The main resource of OST is the engineered language-independent ontology, consisting of a lattice of concepts, each of which is a set of properties (slots, facets, and fillers), including the subsumption and mereological (part-whole) properties that are common to ontologies but adding several hundred other properties, so that the concept HOUSE will look, in its simplified non-proprietary and unfaceted partial property-filler(s) format as follows:

house
    is-a residential building
    has-object-as-part room staircase balcony entrance
    *location street lot square
    *material wood stone metal glass
    *has-object-as-part wall window door foundation roof
    *theme-of build reside

The asterisked properties are inherited from the parent or ancestor concepts. The English lexical items, such as *house*, of course, but also *cottage, villa, mansion, bungalow, cabin* but not *hut, tent, yurt* will be anchored in this concept, as will indeed the Russian *дом, вилла, замок, дворец, дача* be. Processing the sentence *She lived in a big house*, the OST analyzer will read the words, find them in the lexicon, identify the concepts they are anchored in, if any, and try and identify the fillers for the event RESIDE properties (HUMAN and HOUSE will fit into its AGENT and THEME requirements, respectively).

Most of the OST effort is, of course, devoted to the interesting cases of no easy and ideal fit (see the initial sketches for some of those cases in Nirenburg and Raskin 2004, Ch. 8; see Hempelmann *et al.* 2010 for an easy-to-medium case of ambiguity resolved by the technology). The current implementation, even with many necessary modules in the analyzer not yet coded, already analyzes a large chunk of sentences correctly (if you want exact figures read the next section).

Let us now address the standard charges of non-feasibility, subjectivity, and non-scalability raised against us by SML. Our experience and rapid progress towards the product-level implementation, the effort started around 2004, has allowed us to reduce the cost of an individual concept to sligtly over $3 and of the lexical sense to $2.50. We estimate that we need under 15,000 concepts and under 150,000 senses to provide adequate coverage, complemented by the robust unattested input module (see Nirenburg and Raskin 2004: 279–282), already implemented for proper nouns. Doing this from scratch would cost thus $420,000, but much of it is already done and is licensable for less. We are also considering putting up a legacy 6.500-concept ontology and 21,000-sense lexicon (KBAE 2002), after improvement they badly need, as an open source resource. The extension to a new sublanguage/domain, which we have executed 8 times so far, involves, on the average, 6 person months at the post-doctorate level, or under $30,000, and it enriches the ontology by around 50 concepts and the lexicon by around 400 senses (the domains which have hundreds and thousands of terms, such as genomics, bring in many more phrasals but they are highly structured and easy to handle). Finally, translating the senses (not the words!) from one language to another involves a student-level bilingual who does not have to be a linguist. Executed several times fully or partially, it costs under $20,000 per pair of languages, with about $5,000 more for naturalizing the SYN-STRUC zones of the lexical entries in the target language. How does it all compare with multiples of millions spent so far on the SML efforts?

Feasibility involves not just the cost but whether it is at all possible to do it objectively. Well, there is an extreme and useless, even if somewhat plausible view that we all speak idiolects, our own individual subjective languages (see Raskin 1971). So, if you, the reader, understand what we are saying here, it should be reported as a miracle to an appropriate Facebook group! Because, you see, we are using our own idiolect, mysteriously negotiated among the three of us, while you are using your own. The reality, crude as it may be, is that we understand each other most of the time, and, accordingly, the OST acquistion pro-

cess is set up as a hybrid computer-human effort limiting the human performance, basically, to a multiple choice and thus ensuring homogenuity and continuity in the effort. Our elaborate and ever developing acquisition toolbox, increasingly automated, the likes of which would have saved CyC (Lenat 1995) from abandoning its noble initial task of structuring all the «lemmas» of common sense, is briefly sketched out in Taylor *et al.* 2010.

Finally, for the non-scalability charge that should really not be raised from a glass house. Our 15,000-concept and 150,000-sense resources will provide pretty adequate coverage for all possible meaning, this combining the maximum descrptive power with a built-in explanatory power, and our unattested input will do an even better-than-already-implemented job of guessing the senses of unattested input. Our extension to new domains has been tested. What other scalability is there? Oh, language changes, we hear. Indeed, there is that. How much have Russian or English changed in the course of our presentation? Our unattested input, again, can handle the few hundred new words a year, including—most prominently—the English *friend* and *unfriend* as verbs, and the US government prolixity to add several hundred acronyms a month to their special Gobbledegook dialect that does defeat an occasional visitor to Washington, DC, but fortunately, these are forgotten at the same or higher speed. We are developing and perfecting an increasingly automated module for new acquisition , so that unattested input, partially with human approval and correction, be learned, which means that OST includes lexicon learning, and possibly ontology learning (where, incidentally, machine learning techniques could be used, but on TMRs rather than on words and sentences as meaningless character strings.). We are not sure, of course, that full automation will ever be possible, and we refuse to be fazed by it. And how many different training corpora need to be tagged to train the statistical/machine-learning systems for different domains, different corpora, more than two senses per ambiguous word? A friendly question to a co-author of a 2004 workshop presentation at ACL/IJCAI revealed that tagging for annotation, a shallow semantic effort in SML, cost about $75 per sentence per person. Isn't there a scalability/cost issue there?

5. The Evaluation Game That DMA Should Learn From SML (Not Really!)

SML has an enormous advantage over DMA: they have brought the evaluation game to perfection. First, they adjusted it to the very limited, if not only functionality that their technology is capable of, and that they declare the industry/field standard, namely, to identify pertinent documents. The Semeval/Senseval competitions and their informal extensions to other fora (= *forums) give the researchers the quantifiable bragging rights. As Hempelmann and Raskin (2008) polemically claimed, there is definite pride in showing that "our" results outrank "theirs" by .28 %. Early voices objected to these often self-serving metrics as falling far short of real efficacy in product-level applications and user acceptance, and that paucity of real-life applications using SML methods successfully confirms that. The semantic camp is reluctantly developing similar self-encouraging metrics, and we will report them later if we really have to.

The position of the semantic camp has always been that the proof of the pudding is in the eating, not in measuring it according to a number of quantifiable parameters (weight, size, density, color?). The weakness of this position is that the resources have to be developed to a certain minimally functional phase and an application implemented before any informative evaluation. Before it happens, the recall and precision metric is not really applicable. We have been asked whether OST can improve tagging, and our honest response that OST makes tagging unnecessary stuns the younger non-semanticists in NLP who feel that a pillar is being removed from their world. Apparently, the famous Chapaev joke about saddles for the ICBMs is not part of every NLP student's education.

The appropriate evaluation set for the semantic camp should include a number of advanced functionalities that are much harder to develop without understanding the text, such as the paraphrases and similar texts using totally different words that are not from the WordNet synsets; identifying the actual answers to the queries instead of letting users look for them and, more often than not, find them in the documents deemed pertinent; understanding, as humans routinely do, what is left unsaid, namely, inferences, ellipses, implicatures, etc.

With the very low goals set for itself by SML, NLP, like linguistics before it, is getting a very bad name in the field and industries that need our services. The latest disappointed customers are the law firms buying e-discovery products. The yield of the e-mails pertaining to a lawyer's deposition queries is high on recall and very low on precision but the US legislation disallows "fishing expeditions," that is, seeking information on a much broader scope than the case justifies and getting unrelated information. In a growing number of cases, the judge examines the e-discovery yields, sees tons of irrelevant information brought out by the well-pointed but not understood queries, and throws out the entire yield as a "fishing expedition," illegal in US legal system, even if it contains a tiny percentage that is crucial evidence. This leads to lost cases and millions of dollars of damage, and the lawyers are desperate for e-discovery with understanding.

There is an enormous amount of energy and talent on both sides of the semantic divide, and if we set our goals right and channel the energy in that direction we will see an enormous jump in the quality of NLP. We have every reason to believe, as per our skill sets and experience, that the future of NLP depends on the availability of real text understanding. And, incidentally, there is absolutely nothing wrong with statistics or machine learning—as long as we stop applying them to meaningless character strings instead of the elements of meaning, such as ontological property fillers. Until then the competitions will only be marginally meaningful and participation of DMA-type systems not possible, for the simple reason that "AI, linguistics and IR were respectively seeking propositions, sentences and byte-strings and there is no clear commensurability between the criteria for determining the three kinds of entities" (Wilks and Brewster 2009: 47).

6. OST in Action

We have illustrated the OST conceptual apparatus and mode of operation on a simple example of an ambiguous sentence *A dog ate a mouse* in Hempelmann et al. (2010). Now we will demonstrate how OST technology handles a deliberately non-compositional example on, first, *Bill kicked the bucket* and later on *Bill kicked the bucket and dented it*. It should be noted that the Google translation of the latter into Russian arrives at *Билл загнулся и помят его,* thus completely missing the appropriate sense in the first clause.

When the sentence *Bill kicked the bucket* is interpreted by the Semantic Text Analyzer (STAn) with the help of the OST English lexicon and language-independent ontology, the following entries are selected by STAn from the lexicon for consideration:

```
(kick
        [(kick-v1 is domain-dependent and not considered here)]
        (kick-v2
                (cat(v))
                (anno(def "to remove from a place as a result of a violation")
                        (ex "the security kicked the offender out"))
                …
                (syn-struc((subject((root($var1))(cat(np))))
                        (root($var0))(cat(v))
                        (phr((root(out))(cat(phr))))
                        (directobject((root($var2))(cat(np))))))
                (syn-struc1((subject((root($var1))(cat(np))))
                        (root($var0))(cat(v))
                        (directobject((root($var2))(cat(np))))
                        (phr((root(out))(cat(phr))))))
                (sem-struc(remove(precondition(value(^$var99
                                (should-be-a(sem(minor-crime))))))
                        (agent(value(^$var1(should-be-a(sem(human))))))
                        (beneficiary(value(^$var2(should-be-a(sem(human)))))))
        )
        (kick-v3
                (cat(v))
                (anno(def "to punch, usually with the foot")
                        (ex "he kicked the ball. the foot kicked the ball. the wind kicked the ball")
                        (senseprim(1)))
                …
                (syn-struc((subject((root($var1))(cat(np))))
                        (root($var0))(cat(v))
                        (directobject((root($var2))(cat(np))))))
                (sem-struc(kick(agent(value(^$var1)))
                        (instrument(value(^$var1)))
                        (precondition(value(^$var1(should-be-a(sem(physical-event))))))
                        (theme(value(^$var2)))))
        )
        (kick-v4
                (cat(v))
                (anno(def "to die")(ex "the old man kicked the bucket"))
                …
                (syn-struc((subject((root($var1))(cat(np))))
                        (root($var0))(cat(v))
                        (directobject((root(bucket))(cat(np))))))
```

```
                (sem-struc(die(agent(value(^$var1)))))
        )
)
(bucket
        (bucket-n1
                (cat(n))
                (synonyms "")
                (anno(def "a cylindrical vessel or metal, plastic or wood")
                        (ex "he poured water into the bucket"))
                (syn-struc((root($var0))(cat(n))))
                (sem-struc(bucket))
        )
)
(bill
        (bill-n1
                …
                (sem-struc(beak))
        )
        (bill-n2
                …
                (sem-struc(text(has-topic(sem(pay)))))
        )
)
(Bill
        (Bill-pnd1
                …
                (sem-struc(human(gender(value(male)))(has-name(value("Bill")))))
        )
)
```

Next, STAn checks all of the above entries for their mutual compatibility on the basis of the infomration in their SYN-STRUCS, SEM-STRUCS. The SEM-STRUCS are checked against the ontological concept that the entries are anchored in or restricted to (see Raskin et al. 2010 for a formal description of the lexicon; Taylor et al. 2010 and Taylor and Raskin 2010 for a formal description of the ontology and the OST reasoning process). The results of STAn's interpretation of the sentence are:

```
TMR 1:Weight(TMR): 4.24 Event:
                        kick-v3,
                                kick1 agent(value (Bill-pnd1, human1(gender(value(male)))
                                        (has-name(value("Bill"))) ))
                                        theme(value (bucket-n1, bucket1 ))

TMR 2:Weight(TMR): 3.0900002 Event:
                                kick-v4,
                                        die1 agent(value (Bill-pnd1, human1(gender(value(male)))(
                                        has-name(value("Bill"))) ))
```

Notice that STAn recognizes both interpretations of the sentence: Bill died or Bill hit a physical object. Now let us consider the sentence *Bill kicked the bucket and dented it*.

```
        (dent-v1
                …
                (cat(v))
                (syn-struc((subject((root($var1))(cat(np))))
                        (root($var0))(cat(v))
                        (directobject((root($var2))(cat(np))))))
                (sem-struc(damage(relative-force(less-than(0.3)))
                        (agent(value(^$var1)))
                        (theme(value(^$var2)))))
        )
```

There are several modules that will be activated to process this sentence, in addition to the previously mentioned ones. The sense of *dent* requires a subject and an object, the subject will be selected to be *Bill*, and the object, the pronoun *it*. The next step is to resolve the pronoun (Co-Reference Module). There is only one satisfactory candidate, and that is the concept that corresponds to the word *bucket*. Notice that while such a concept exists in TMR1, the interpretation of the idiomatic expression in TMR2 removes such possibility. Thus, the combination of the two clauses in the new sentence is only possible with TMR1. This successfully disambiguates *kick the bucket* and removes the sense of dying or загнулся from the table. When STAn's Event Embedment Module checks the possible relationship between KICK and DAMAGE in the ontology, we will also find that DAMAGE is an effect of KICK, resulting in the following interpretation of the sentence:

```
kick1    (agent(value (human1
               (gender(value(male)))
               (has-name(value("Bill")))
         )))
         (theme(value (bucket1)))
         (effect(value(damage
               (relative-force(less-than(0.3)))
               (theme(value(bucket1)))
         )))
```

While this is a constructed example, it demonstrates the capability of the system on the actual, real-life ontology and lexicon of an implemented system. It is this capability that keeps all interpretations of the ambiguous sentences when needed, and removes them when there is enough knowledge to provide accurate results in machine understanding of natural language.

7. Conclusion: Crawling, Flying, and Other Self-Propelling Modes

At a major NLP gathering a few years into this century, a fellow workshop participant, a master grant getter, declared, in fake admiration, that the semantic approach was about flying while he and his non-semantic confrères were crawling, which he made sound as something in hand (his ambitious project has yielded no known result). Maxim Gorky also had something to say about crawling and flying but he never got a single Federal grant... We think it is actually neither about crawling nor flying but rather about walking, preferably running, with one's feet firmly on the ground (most of the time) and the pace rapidly accelerating, to the goal (not Grail) that the industrial, societal and academic needs are making it increasingly urgent for us to reach. One cannot help recalling here this once famous (Hubert) Dreyfus (1992: 100) quote: "...the first man to climb a tree could claim tangible progress toward reaching the moon. Rather than climbing blindly, it's better to look where one is going." Are we, the computational semanticists, just a leg or so ahead because we are already trying? Should we really be spending enormous amounts of money and effort on just divining what is hidden behind the door of meaning that we presume closed, or should we continue the difficult but, ultimately, less costly work on the doorstop that keeps moving the door towards the wall, wider and wider, even if possibly we will never succeed to take it off the hinges? Will the curiosity kill the cat? Meow!

References

1. *Agirre E., Marquez L. and Wicentowski R. (eds.). Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval)//Association for Computational Linguistics, Prague, Czech Republic, 2007.*
2. *Ахманова О. С., Мельчук И. А., Падучева Е. В. и Фрумкина Р. М.* О точных методах исследования языка.//М.: Издательство Московского государственного университета, 1961.
3. *Bar-Hillel Y.* A demonstration of the non-feasibility of fully automatic high quality translation, 1954. Reprinted in his: Language and Information //Jerusalem: Magness, 1960.
4. *Chomsky N.* Aspects of the Theory of Syntax // Cambridge, MA: MIT Press, 1965.
5. *Dreyfus H. L.* What Computers Still Can't Do: A Critique of Artificial Reason.//Cambridge, MA: MIT Press, 1992
6. *Firth J. R.* Modes of meaning. In his: Papers in Linguistics.//London: Oxford University Press, 1957, P. 19–215.
7. *Frege G.* Die Grundlagen der Arithmetik. Eine logisch-mathematische Untersuchung ueber den Begriff der Zahl.// Breslau: Koebner., 1884. English translation by J. L. Austin, Foundations of Arithmetic.//Oxford: Blackwell, 1950.
8. *Gale W. A., Church K. and Yarowski D.* A method for disambiguating word senses in a corpus.// Comput. Human, 26: P. 415–439, 1992.
9. *Hempelmann C. F., Raskin V. and Taylor J. M.* In: Proceedings of ICAI '10.//Las Vegas, NE, July 2010.
10. *Hempelmann C. F. and Raskin V.* Semantic search: Content Vs. formalism. //Proceedings of LangTech 2008, http://www.langtech.it/en/technical_program/technical_program.htm, 2008.
11. *Hjelmslev L.* Prolegomena to a Theory of Language.// International Journal of American Linguistics, Suppl. to vol. XIX, 1953.
12. *Hjelmslev L.* Dans quelle mésure les significations des mots peuvent-elles etre considérées comme formant une structure?//Proceedings of the Eighth International Congress of Linguists, Oslo, 636–654, 1958.

13. *Ide N. and Veronis J.* Word sense disambiguation: The state of the art.//Computat. Ling. *24,* 1: 1–40, 1998.
14. *Ide N. and Wilks Y.* Making sense about sense. // In: E. Agirre and P. Edmonds (eds.), Word Sense Disambiguation: Algorithms and Applications.// Springer, New York, NY, 47–73, 2006.
15. *Katz J. J. and Fodor J. A.* The structure of a semantic theory.//Language 39(1): 170–210, 1963.
16. *KBAE:* Knowledge-Based Acquisition Editor. Purdue Version 2.1. http: //kbae.cerias.purdue.edu:443/; guest login (browsing only), nlpgroup; password, ch@ng3me.//NLP Lab and CERIAS, Purdue University W. Lafayette, IN, 2002.
17. *Kilgariff A.* I don't believe in word senses // Comput. Human. *31,* 2, 91–113, 1997.
18. *Kilgariff A.* Senseval: An exercise in evaluating word sense disambiguation programs?? LREC, Granada, Spain, 1255–1258, 1998.
19. *Kilgariff A.* Word senses. // In: E. Agirre and P. Edmonds (eds.), Word Sense Disambiguation: Algorithms and Applications // Springer, New York, 29–46, 2006.
20. *Languages and Machines.* Computers in Translation and Linguistics. A Report by Automatic Language Processing Advisory Committee, Division of Behavioral Studies, National Academy of Sciences, National Research Council, Publication 1416 // Washington, DC, 1966.
21. *Lenat D. B.* Cyc: A large-scale investment in knowledge infrastructure // Communications of the ACM 38(11), 1995.
22. *Mallery J. C.* Thinking about foreign policy: Finding an appropriate role for artificial intelligence computers.//Unpublished Ph. D. dissertation. Political Science Department, MIT, Cambridge, MA, 1988.
23. *Manning C. D. and Schuetze H.* Foundations of Statistical Natural Language processing. Second Printing with Corrections.//Cambridge, MA: MIT Press, 2000.
24. *McCawley J. D.* Everything That Linguists Have Always Wanted to Know About Logic but Were Ashamed to Ask, 2nd ed. // Chicago: University of Chicago Press, 1993.
25. *Mitchell T. M.* Machine Learning // New York: McGraw Hill, 1997.
26. *Navigli R.* Word sense disambiguation: A survey.// ACM Comput. Surv. 41, 2, Article 10 (February 2009), http://doi.acm.org/10.1145/1459352.1459355, 2009.
27. *Ng T. H.* Getting serious about word sense disambiguation.//In: Proceedings of the ACL SIGLEX Workshop on Tagging Text with Lexical Semantics: Why, What and How? (Washington D.C.). 1–7, 1997.
28. *Nirenburg S. and Raskin V.* Ontological Semantics.//Cambridge, MA: MIT Press, 2004.
29. *Peirce C. S.* On the Nature of Signs. In: J. Hoopes (ed.), Peirce on Signs : Writings on Semiotic by Charles Sanders Pierce // Chapel Hill, NC: University of North Carolina Press, 116–141, 1991.
30. *Petrenko M.* Ontological semantics and abduction: parsing ellipsis. // Компьютерная лингвистика и интеллектуальные технологии, по материалам ежегодной Международной конференции "Диалог 2009." Выпуск 8 (15). /Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue 2009."Issue 8 (15). Moscow: RGGU Press, 2009.
31. *Раскин В.* К теории языковых подсистем///М.: Издательство Московского государственного университета, 1971. Второе дополненное издание.//Москва: Издательство ЛКИ, 2007.
32. *Raskin V.* Theory and practice of justification in linguistics.//In: P. L. Clyne et al. (eds.), Papers from the Parasession on the Elements, Chicago: Chicago Linguistics Society, 1979.
33. *Raskin V.* Fear of Semantics. Invited lecture.//Center for Machine Translation, Carnegie Mellon University, Pittsburgh, 1988.
34. *Raskin V.* Frawley: Linguistic Semantics.//Language 70(3): 552–556, 1994.
35. *Raskin V.* The whys and hows of ontological semantics. // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции "Диалог 2006" (Бекасово, 31 мая — 4 июня 2006 г.). / Под ред. Н. И. Лауфер, А. С. Нариньяни, В. П. Селегея / N. I. Laufer, A. S. Narinyani and V. P. Selegei (eds.), Computational Linguistics and Intellectual Technologies. International Conference "Dialogue 2006" Proceedings (Bekasovo, 31 May — 4 June, 2006)/. Moscow: RGGU Press, 2006.
36. *Raskin V., Taylor J. M. and Hempelmann C. F.* Ontological Semantic Technology for Detecting Insider Threat and Social Engineering.//Submitted to NSPW 2010
37. *Saussure F. de.* Cours de linguistique générale, ed. C. Bally and A. Sechehaye, with the collaboration of A. Riedlinger.//Lausanne and Paris: Payot 1916. English translation by W. Baskin, Course in General Linguistics.//Glasgow: Fontana/Collins, 1977.
38. *Шайкевич А. Я.* Распределение слов в тексте и выделение семантических полей. //Иностранные языки в школе. Вып. 2. М., 1963.
39. *Taylor J. M. and Raskin V. Fuzzy Ontology for Natural Language.//Submitted to NAFIPS 2010,*
40. *Taylor J. M., Hempelmann C. F. and Raskin. V.* On an automatic acquisition toolbox for ontologies and lexicons.// Proceedings of ICAI '10.//Las Vegas, NE, July 2010.
41. *Trier J.* Der Deutsche Wortschatz im Sinnbezirk des Verstandes.//Heidelberg: Winter, 1931.
42. *Turing A. M.* Computing machinery and intelligence.//Mind *54*, 443–460, 1950.
43. *Weisgerber J. L.* Das Gesetz der Sprache als Grundlage des Sprachstudiums.//Heidelberg: Quelle und Meyer, 1951.
44. *Wilks Y. and Brewster C.* Natural Language Processing as a Foundation of the Semantic Web. *Foundations in Web Science 1/3–4,* P. 199–327, 2006.

Гармонизация систем помет для многоязычных корпусов посредством решетки понятий

Harmonizing tagsets for multilingual corpora via concept lattice*

Rosen A. (alexandr.rosen@ff.cuni.cz)

Institute of Theoretical and Computational Linguistics, Faculty of Arts, Charles University, Prague, Czech Republic

Сравнение систем морфосинтаксических помет обнаруживает различные предположения, заслоняющие сходства и различия между языками. Чтобы преодолеть формальные и концептуальные несоответствия, мы строим абстрактную межъязыковую систему помет как иерархию категорий, используя анализ формальных понятий.

1. Introduction

Multilingual corpora can be annotated with morphosyntactic tags by monolingual tools. However, each of the tools is typically bundled with a specific tagset. This variety of tagging schemes may be a problem for the user: InterCorp, a parallel corpus, currently offers on-line concordances in 22 languages, 11 of them tagged with 11 different tagsets[1]. Fig. 1 illustrates the tagset variety using comparable examples of prepositional phrases in all of the 11 presently tagged languages[2].

We are aiming at a solution that would delegate the task of dealing with multiple tagsets to the system, allowing the user to interact with an abstract interlingual hierarchy of linguistic categories, a common "tagset" that is only used for mediating between language-specific tagsets, not to tag real texts. In order to reflect the differences between various tagsets, the common "tagset" takes three different perspectives of word class. Thus, the tag for the Czech relative pronoun *který* 'which' is decoded as a category with the properties of lexical pronoun, inflectional adjective and syntactic noun, each with its appropriate morphological characteristics.

Tags in all tagsets can be described as objects with properties and the methods of Formal Concept Analysis [2] can be used to construct the hierarchy automatically as a concept lattice and to (partially) resolve tag queries that do not quite match the tags used for the specific language, in a way similar to that employed by Janssen [3] for dealing with lexical gaps in a multilingual lexical database.

This is certainly not the first attempt to design an interlingual representation of linguistic categories in the context of multilingual corpora. We wish to mention at least *MULTEXT-East* [4], whose tagging scheme became a *de facto* standard for inflectional languages, and *Interset*, a truly interlingual tagset [5], designed primarily for translating tags from one tagset into another. However, neither quite satisfies our requirements: they miss some categorial correspondences between languages and do not support the idea of arbitrary levels of specificity (see fig. 1).

2. Word classes in three flavours

The traditional list of eight word classes is defined by a mix of morphological, syntactic and semantic criteria. For nouns or adjectives the three criteria agree. Nouns decline independently in typical nominal positions, referring to entities; attributive or predicative adjectives, representing properties, agree with nouns. On the other hand, numerals and pronouns are defined solely by semantic criteria, while their syntactic and morphological behaviour is rather like that of nouns (cardinals and personal pronouns) or adjectives (ordinals and possessive pronouns). For such cases, the option of abandoning the traditional list in favour of a cross-

* Work on this project was supported by grant no. MSM0021620823 of the Czech Ministry of Education, Youth and Sports.

[1] For more details about the project see [1] or the project site at http://korpus.cz/intercorp/. The corpus can be queried at korpus.cz/Park after registration at http://ucnk.ff.cuni.cz/english/dohody.php.

[2] For details about the tagging tools and tagsets see http://korpus.cz/english/intercorp-info.php. Here and below, Czech positional tags are truncated: **RR-6** stands for **RR-6----------** (tag for a preposition selecting local case).

en	in	the	remotest	exurbs
	IN	**DT**	**JJS**	**NNS**
de	in	den	abgelegensten	Außenbezirken
	APPR	**ART**	**ADJA**	**NN**
nl	in	dit	schitterende	appartement
	600	**370**	**103**	**000**
fr	dans	les	plus lointaines	banlieues
	PRP	**DET:ART**	**ADV ADJ**	**NOM**
sp	en	las	zonas	más remotas
	PREP	**ART**	**NC**	**ADV ADJ**
it	da	queste	lingue	babeliche
	PRE	**PRO:demo**	**NOM**	**ADJ**
ru	v	samych	otdaljonnych	rajonach
	Sp-l	**P—pl**	**Afp-plf**	**Ncmpln**
cs	v	těch	nejodlehlejších	zástavbách
	RR–6	**PDXP6**	**AAFP6----3A**	**NNFP6-----A**
bg	na	tova	prijatelsko	dviženie
	R	**Pde-os-n**	**Ansi**	**Ncnsi**
pl	w	tym	wspaniałym	apartamencie
	prep:loc:nwok	**adj:sg:loc:m3:pos**	**adj:sg:loc:m3:pos**	**subst:sg:loc:m3**
hu	a	szép	katalán	lányba
	ART	**ADJ**	**ADJ**	**NOUN(CAS(ILL)**

Figure 1. Differences in tagging: prepositional phrases

classification along the three dimensions seems attractive. Distinctions between the three aspects are borne out also by the tagsets. Our tagset for Czech has a preference for lexically-based classification, the Polish tagset for inflectional word classes, the German tagset distinguishes pronouns by their syntactic function.

Fig. 2 shows a simple case — nouns and adjectives are nouns and adjectives, respectively, on all three criteria.3 The topmost node *wcl* stands for both nouns and the adjectives. Its daughters are labelled by the three aspects: *lexical* (for 'semantic'), *inflectional* (for 'morphological') and *syntactic*.4 The boxes around the labels suggest that the sets of objects denoted by the nodes have a non-empty intersection. In fact, all four sets involved are identical, which is a feature of cross-classification. The other nodes stand for word classes in the three respective flavours, distinguished in their labels by the initial letter. The six types of word classes share only two daughters, the objects to be classified. Each of the two objects inherits the property of being a word class according to the three criteria.

The hierarchy of categories or *types* is partially ordered by their specificity. Each type denotes a set of objects — language-specific tags, identified by their name and specific tagset. The topmost type denotes all tags in all tagsets. Immediate subtypes of a supertype denote subsets of that supertype. A tag in the denotation of the supertype must be in the denotation of at least one of the subtypes. A subtype can have more than one supertype. In this case, the subtype denotes a subset of the intersection of the sets denoted by its supertypes.

Unlike regular nouns and adjectives, a Czech *wh*-form *který* 'which' in its use as a relative (rather than interrogative) pronoun belongs to three different word classes at the same time. In (1), *který* is at the same time a *syntactic* noun as the subject of the relative clause, a *lexical* pronoun with "dog" as its antecedent, and — due to its adjectival declension — an *inflectional* adjective (see fig. 2).

(121)

Psa, který nemá náhubek, do vlaku nepustí.
dog_{ACC} $\text{which}_{\text{NOM}}$ has_{NEG} $\text{muzzle}_{\text{ACC}}$ into train $\text{let in}_{\text{NEG,PL,3RD}}$
'An unmuzzled dog won't be allowed on the train.'

To express this triple membership, the Czech tag **P4** for relative pronouns[5] is a subtype of the cross-classifying word classes, each representing a different dimension — see fig. 3.

The fragment can be extended by other objects as in fig. 4: cardinal and ordinal numerals, personal, possessive and interrogative pronouns. Ordinals such as *pátý* 'fifth' are treated as *lexical* numeral and adjective — both *inflectional* and *syntactic*. Possessive

[3] All hierarchies shown here are partial: they cover only a fraction of morphological categories and languages.

[4] We use *lexical* rather than *semantic* — *lexical* word classes have their properties specified in the lexicon.

[5] We ignore all but the first two positions in the tag.

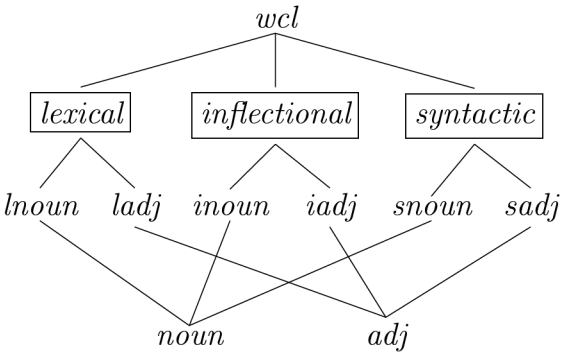


Figure 2. Nouns and adjectives are nouns and adjectives from all three aspects

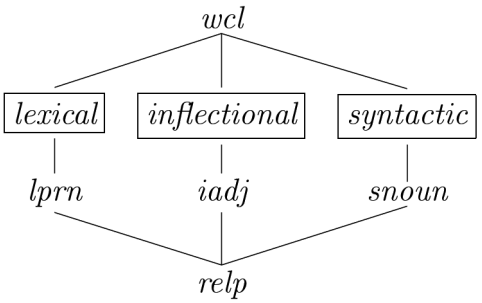


Figure 3. A hierarchy fragment for the Czech relative pronoun ***který*** 'which'

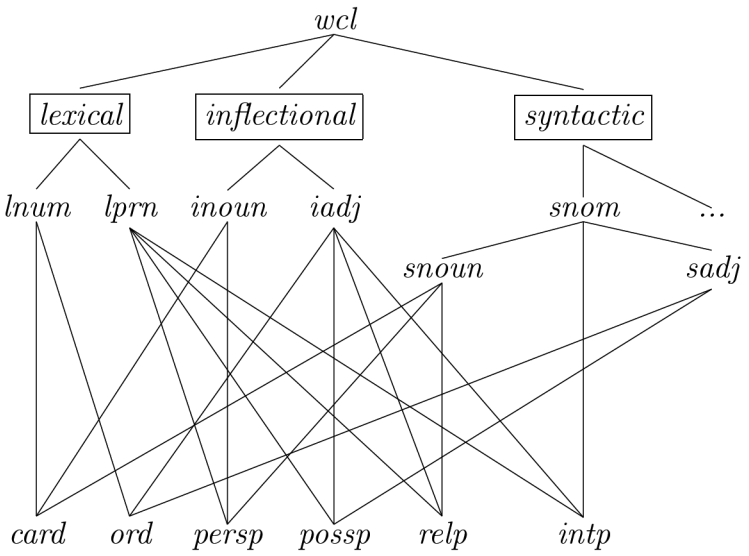


Figure 4. Distinguishing types of numerals and pronouns in a hierarchy

pronouns differ in being *lexical* pronouns. Personal pronouns are inflectional and syntactic nouns, similarly as cardinal numerals. The interrogative homonym of *který* in its relative use can be used as a syntactic adjective or noun. The node *intp* inherits from *snom*, representing syntactic nouns *or* adjectives, while *relp* can only be a syntactic noun, due to its ancestor *snoun*.

However, there is a single Czech tag covering both the relative and interrogative use of *který* (**P4**), which should be represented as ambiguous between relative pronoun and syntactic noun on the one hand and interrogative pronoun and syntactic adjective or noun on the other. The modified hierarchy in fig. 5 captures this ambiguity. The Czech tag **P4** corresponds to a node labelled $lprn \wedge iadj \wedge snom$.

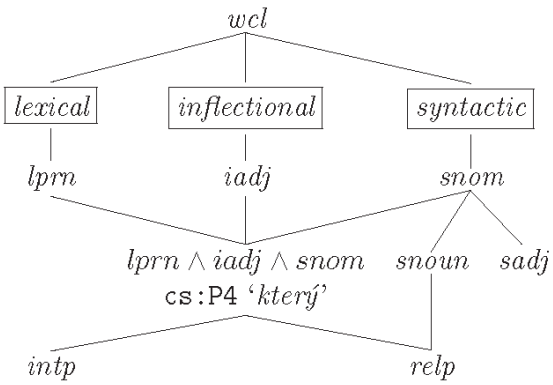


Figure 5. A single node for interrogative and relative pronouns

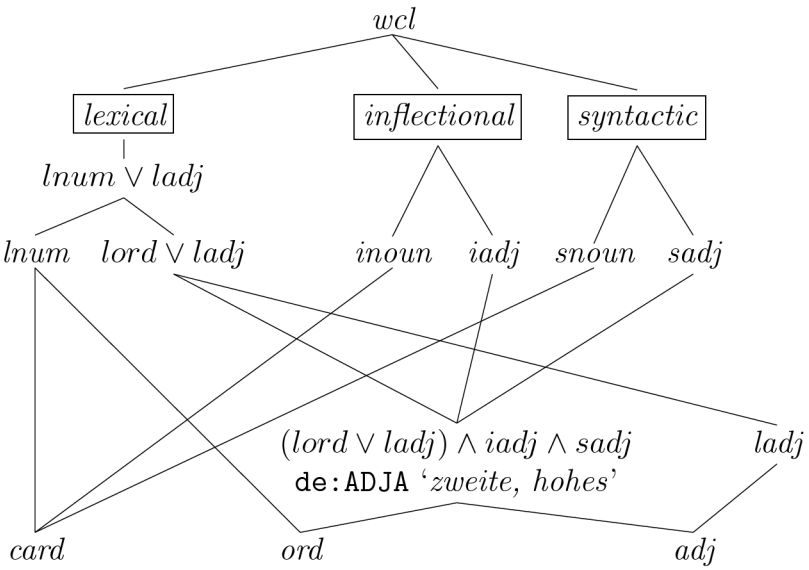


Figure 6. A single node for ordinal numerals and adjectives

The three views of word class allow for proper mapping between language-specific tagsets. The tag for adjective in the English, German, French, Italian and Polish tagsets covers also ordinal numerals. If all these tags are mapped as *syntactic* adjectives, they end up correctly in the same class as Czech, Spanish, Russian or Bulgarian adjectives, ordinal numerals and possessive pronouns. Their *lexical* word class is unknown, although it is not arbitrary. Fig. 6 shows a fragment of the hierarchy with a node representing both ordinal numerals and adjectives, labelled (*lord* ∨ *ladj*) ∧ *iad j* ∧ *sadj* and corresponding to the German tag **ADJA**.

The German ordinal number *zweite*, tagged as adjective (similarly as *hohes*), is a subtype of inflectional and syntactic adjective (*iadj* and *sadj*), and also a subtype of a general type covering lexical adjectives and ordinal numerals (*ladj* ∨ *lord*).

Partial hierarchies can be merged. The result of merging the above two hierarchies (figures 5 and 6) is shown in fig. 7.

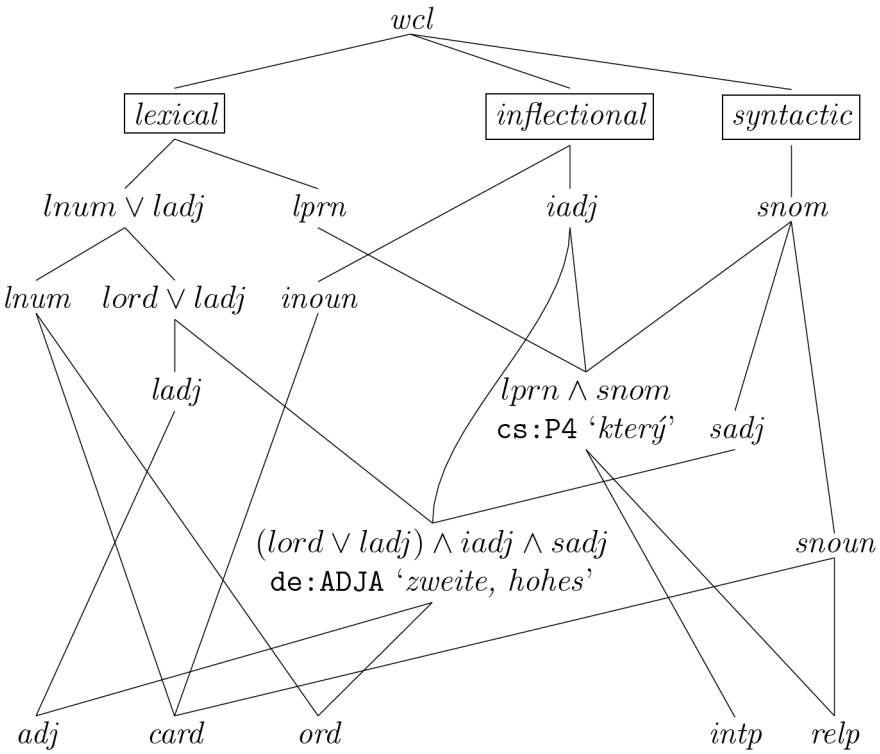


Figure 7. Hierarchies in figures 5 and 6 merged

We have barely scratched the surface of the topic of cross-classifying word classes. Obvious candidates for this treatment could be derived words. However, the possibility of multiple derivation and the constraints of the language-specific tagsets may present a prohibitive obstacle to any significant extension of the approach.

3. Morphological categories

Tags often encode more information than just word class. Word class of any flavour may be required to co-occur with a set of other categories: personal and possessive pronouns with the *lexical* categories of person, number and gender, inflectional adjectives with the *inflectional* categories of gender, number and case. A possessive pronoun such as *jejího* is *lexically* 3rd person, singular and feminine, while *inflectionally* it is masculine or neuter, singular, genitive or accusative (2).6

(2) *Martina je moje sousedka.*
Martina is my neighbour$_{\text{FEM,SG,NOM}}$.
Jejího syna často potkávám v tramvaji.
her$_{\substack{\textit{lex}\text{: 3RD,FEM,SG;}\\ \textit{infl}\text{: MASC,SG,ACC}}}$ son$_{\text{MASC,SG,ACC}}$ often meet$_{\text{1ST,SG}}$ in tram.

'Martina is my neighbour. I often meet her son on the tram.'

The set of categories appropriate to a word class may be defined as types in the hierarchy, which further cross-classify types corresponding to language-specific tags. Then the user can refer to all plural items by specifying them merely as **pl**.

The tag for the Czech possessive pronoun *jejího* in fig. 8 is a subtype of lexical pronoun (*lprn*) and inflectional adjective (*iadj*).7 As a possessive pronoun, it is required by the hierarchy8 to be a subtype of *lexical* gender (*lgend*), number (*lnum*) and person (*lpers*), more precisely of their intermediate subtypes, specifying morphological categories. As an inflectional adjective, it is required to be a subtype of *inflectional* gender *igend*, case (*icase*) and number (*inum*). In isolation, the form *jejího* is ambiguous between (inflectional) genitive and accusative and inflectional masculine and neuter genders. As the tag suggests, the former ambiguity is assumed to be resolved (the digit "4" at the 5th position stands for accusative), unlike the latter ambiguity, which is retained (the character "Z" at the third position stands for all genders, except feminine). Therefore, the tag is a subtype of *imasc* ∨ *ineut*, covering both *imasc* and *ineut*.

6 Czech personal and possessive pronouns share the same *lexical* categories and are distinguished by their *inflectional* category.

7 It is also a subtype of syntactic adjective. Types less relevant for the current discussion are omitted for brevity.

8 More general co-occurrence restrictions could be specified at a meta-level to ease the initial manual task of mapping tags to categories.

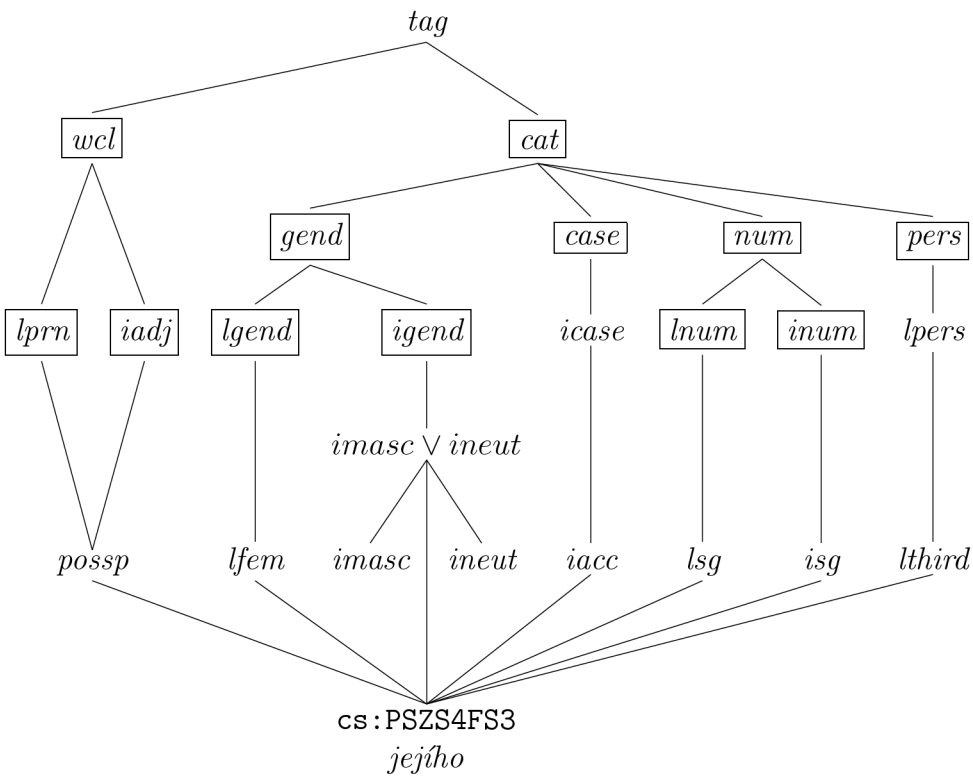


Figure 8. Morphological categories used to tag a Czech possessive pronoun jejího, a category-based view

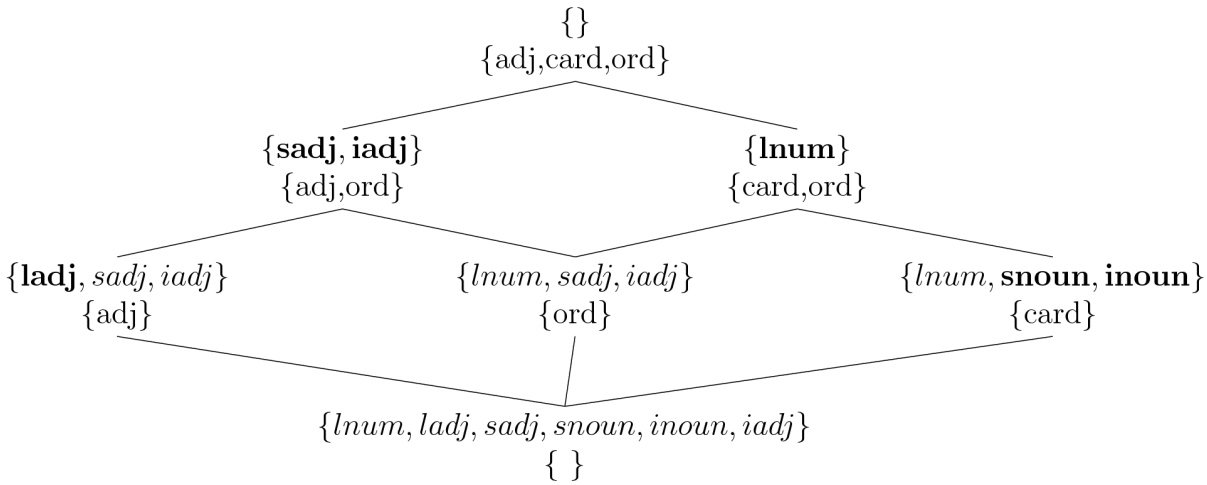


Figure 9. Morphological categories used to tag a Czech possessive pronoun jejího, a lexical/inflectional view

The hierarchy in fig. 8 leaves the lexical/inflectional distinction implicit. In fig. 9 this distictions is shown at the top level, as in all previous hierarchies. For clarity, general category labels (*gend*, *case*, etc.) are omitted.

4. Building and using the common tagset

The type hierarchies presented so far are similar to concept lattices of Formal Concept Analysis (FCA), a logical formalism equipped with methods for constructing and using the lattices [2,6]. The task of FCA is to classify objects according to their properties (attributes). The classification is based on the notion of *concept*, consisting of a set of objects as its extension and a set of attributes as its intension.

The first step of the analysis is to identify the objects and their attributes. This is done in a tabular data structure called *formal context*. Table 1 is an example of a formal context for our previous example of adjectives and numerals (fig. 6). Attributes corresponding to the boxed labels in fig. 6 are omitted: they would be specified for all objects and would not make the resulting lattice more informative.

Table 1. Formal context for adjectives and ordinal numerals

	ladj	*lnum*	*iadj*	*inoun*	*sadj*	*snoun*
adj	•		•		•	
ord		•	•		•	
card		•		•		•

Next, a set of formal concepts is built, each of the concepts consisting of a pair of the set of objects, and a set of attributes. Objects belonging to a concept belong also to its superconcept and the concepts are partially ordered by specificity (roughly: the more attributes, the more specific).

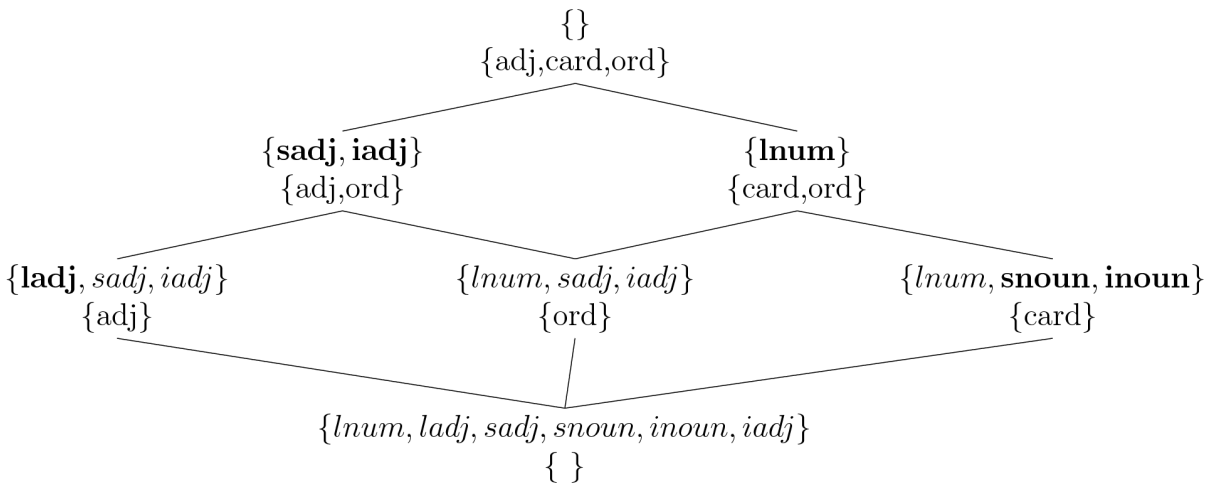


Figure 10. Concept lattice for adjectives and ordinal numerals

Table 2. Formal concepts derived from table 11

1	⟨{adj,ord,card},	{}⟩
2	⟨{ord,card},	{*lnum*}⟩
2	⟨{adj,ord},	{*iadj,sadj*}⟩
3	⟨{adj},	{*ladj,iadj,sadj*}⟩
3	⟨{ord},	{*lnum,iadj,sadj*}⟩
3	⟨{card},	{*lnum,inoun,snoun*}⟩
4	⟨{},	{*ladj,lnum,iadj,inoun,sadj,snoun*}⟩

Finally, the concept lattice can be drawn (fig. 10). Its geometry is significantly simpler than the hierarchy constructed intuitively (as in fig. 6), while the concept ambiguous between adjectives and cardinal numerals is still present. The last two steps can be done automatically.[9]

The concept lattice can be used for reasoning about attributes, as in the following implications: *ladj* ⇒ *sadj* or *snoun* ⇒ *lnum*. Such statements can be used to assist the user in making queries including language-independent category labels (such as "adj"), or to match incompatible language-specific tags.

The concept with the extension {ord} corresponds to **cs:Nr**, the Czech tag for ordinal numerals, while the concept with the extension {adj,ord} corresponds to **de:ADJA**, the German tag covering adjectives and ordinal numerals. To look up its Czech equivalent we have to find a Czech tag corresponding to the {adj,ord} concept. In the absence of such a tag, the more specific concepts are traversed and the disjuction of Czech tags corresponding to {adj} and {ord} is the result. Looking up a German equivalent of **cs:Nr** is similar to the scenario when the user asks for "ord" in a German text. It's easy in a Czech text, because the appropriate tag **cs:Nr** is available. For German, there is no tag corresponding to "ord". There are also no concepts more specific than {ord} that would correspond to German tags. The only option is to resort to a more general concept {adj,ord}, with a corresponding German tag. The extensions of the two concepts can be compared and the user warned that she would have to filter out concordances including categories corresponding to "adj".

Attributes specified for an object in a formal context are interpreted in conjuction. Thus, specifying both *snoun* and *sadj* as attributes of an interrogative pronoun (intp) would mean that it is simultaneously syntactic noun and a syntactic adjective. To model disjunction of attributes we have to introduce a more general attribute covering the two options. The formal context and concepts for numerals and pronouns are shown below in tables 3 and 4 and the corresponding lattice in fig. 11.

This is not the first application of FCA in the field of linguistics, not even in a multilingual setting. Priss [7] gives an overview of linguistic applications of FCA and Janssen [3] is concerned with multilingual lexical databases. His lattice, a structured lexical interlingua connecting words from different languages, is similar to the common abstract tagset. Given that the world of morphosyntactic tags is simpler than the world of words, this is a reassuring finding, reinforced by the continuing advances of FCA and its application to other very complex domains.

Table 3. Formal context for numerals and pronouns

	lnum	*lprn*	*inoun*	*iadj*	*snoun*	*sadj*	*snom*
card	•		•		•		•
ord	•			•		•	•
persp		•	•		•		•
possp		•		•		•	•
relp		•		•	•		•
intp		•		•			•

Table 4. Formal concepts derived from table 3

1	⟨{card,ord,persp,possp,relp,intp},	{snom}⟩
2	⟨{card,ord},	{lnum,snom}⟩
2	⟨{card,persp,relp},	{snoun,snom}⟩
2	⟨{ord,possp,relp,intp},	{iadj,snom}⟩
2	⟨{persp,possp,relp,intp},	{lprn,snom}⟩
3	⟨{card,persp},	{inoun,snoun,snom}⟩
3	⟨{ord,possp},	{iadj,sadj,snom}⟩
3	⟨{persp,relp},	{lprn,snoun,snom}⟩
3	⟨{possp,relp,intp},	{lprn,iadj,snom}⟩
4	⟨{card},	{lnum,inoun,snoun,snom}⟩
4	⟨{ord},	{lnum,iadj,sadj,snom}⟩
4	⟨{persp},	{lprn,inoun,snoun,snom}⟩
4	⟨{possp},	{lprn,iadj,sadj,snom}⟩
4	⟨{relp},	{lprn,iadj,snoun,snom}⟩
5	⟨{},	{lnum,lprn,inoun,iadj,snoun,sadj,snom}⟩

[9] See http://www.fcahome.org.uk/fca.html.

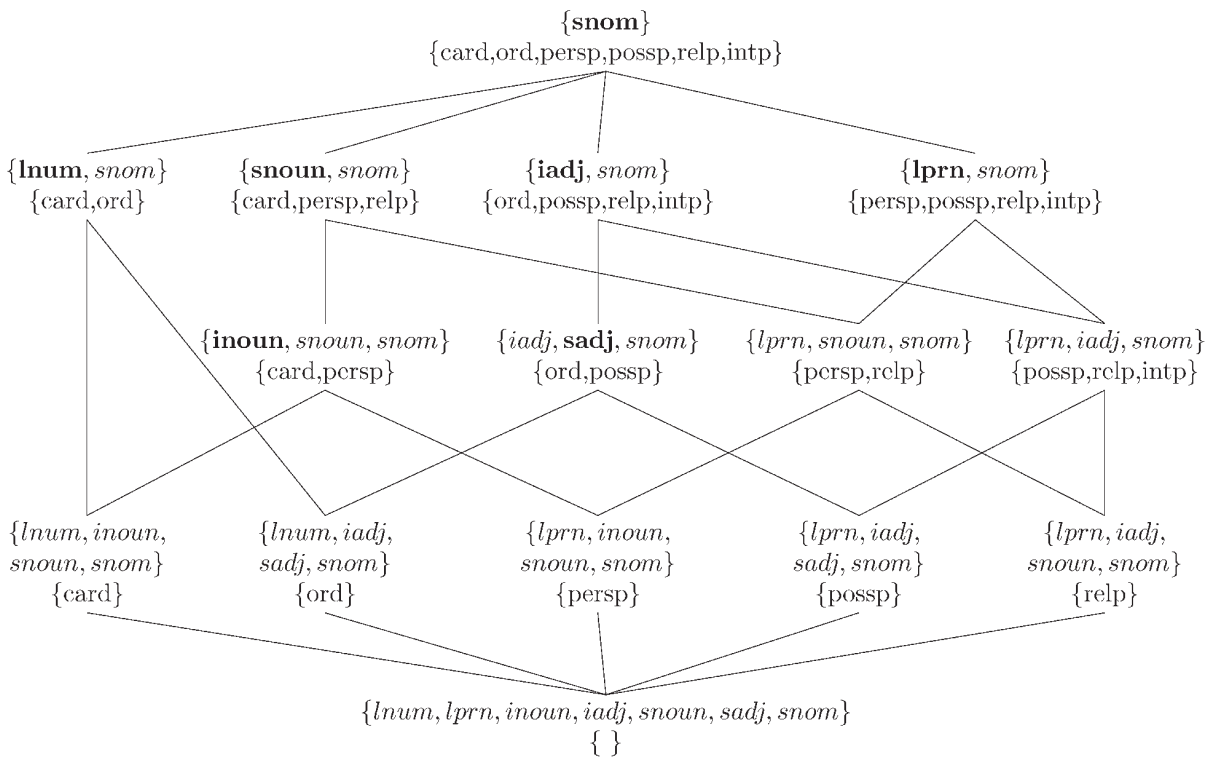


Figure 11. Concept lattice for numerals and pronouns

5. Conclusion

A solution to the problem of tagset variety in a multilingual corpus can be an abstract, hierarchically structured interlingual tagset, based on a three-way distinction in the system of word classes, allowing for intuitive and underspecified queries and principled mappings between different language-specific tagsets. If corpus data include only original, language-specific tags, the system can be easily modified and extended without touching the corpus data and the abstract categories can be mapped to tags in any format.

The cost is higher complexity, both conceptual and formal/implementational: a module to resolve queries using the type hierarchy specification is needed. However, we believe that the price is well justified and that the modular framework of our proposal allows for customising the setup of the system according to specific preferences. Formal Concept Analysis seems to be the answer to concerns about the costs of designing the hierarchy.

Obviously, more work is needed: although some mapping to language-specific tagsets can be acquired from existing resources such as *Interset*, specifying formal contexts in FCA is tedious even for a single language, even more so without the options of multi-valued attributes, disjunctive values and co-occurrence restrictions, all waiting to become part of the system, together with interfaces to concordancers and other applications.

References

1. *Vavřín M., Rosen A.* InterCorp: A Multilingual Parallel Corpus Project // Proceedings of the International Conference Corpus Linguistics — 2008. St. Petersburg State University, 2008. P. 97–104
2. *Ganter B., Wille R.* Formal Concept Analysis. Mathematical Foundations // Berlin/Heidelberg: Springer, 1999.
3. *Janssen M.* Multilingual Lexical Databases, Lexical Gaps, and SIMuLLDA // International Journal of Lexicography, 2004. 17 № 2.
4. *Erjavec T.* Harmonised Morphosyntactic Tagging for Seven Languages and Orwell's 1984 // Proceedings of the 6th Natural Language Processing Pacific Rim Symposium, NLPRS'01, 2001. P. 481–492.
5. *Zeman D.* Reusable tagset conversion using tagset drivers // Proceedings of the Language Resources and Evaluation Conference, LREC 2008, Marrakech, Marocco, 2008.
6. *Wille R.* Formal concept analysis as mathematical theory of concepts and concept hierarchies // Ganter B., editor, Formal Concept Analysis. Foundations and Applications, volume 3626 of Lecture Notes in Artificial Intelligence, Berlin/Heidelberg: Springer, 2005. P. 1–33.
7. *Priss U.* Linguistic applications of formal concept analysis // Ganter B., editor, Formal Concept Analysis. Foundations and Applications, volume 3626 of Lecture Notes in Artificial Intelligence, Berlin/Heidelberg: Springer, 2005. P. 149–160.

Using a domain ontology for the semantic-statistical classification of specialist hypertexts

Noah Bubenhofer (bubenhofer@ids-mannheim.de),
Roman Schneider (schneider@ids-mannheim.de)

Institute for German Language (IDS), Mannheim/Germany

In this feasibility study we aim at contributing at the practical use of domain ontologies for hypertext classification by introducing an algorithm generating potential keywords. The algorithm uses structural markup information and lemmatized word lists as well as a domain ontology on linguistics. We present the calculation and ranking of keyword candidates based on ontology relationships, word position, frequency information, and statistical significance as evidenced by log-likelihood tests. Finally, the results of our machine-driven classification are validated empirically against manually assigned keywords.

1. Introduction

Research into ontologies has received much attention for the last years [16] [17] [18]. Due to its practical use for common tasks related to knowledge sharing and publication, it has been subject of study in most different scientific communities. Ontologies are often seen as enabling technology for information sharing, with their ability to be easily reused being a key factor for successful application scenarios [4] [6] [8] [15]. On the web, which represents a large universe of mostly unclassified semi-structured hypertexts, semantic techniques and technologies open up new strategies for information retrieval and text classification [5].

The Institute for German Language (IDS) in Mannheim is the central institution for research and documentation of the German language. It hosts several specialist resources, including the hypertextual information systems Grammis and ProGr@mm and a terminological ontology [12] [13] [14]. Since only less than 40 % of the hypertexts are classified with manually assigned keywords, our goal is to gain insight of how ontology features can affect automatic semantic-statistical classification. We introduce the resources as far as necessary to understand our test-bed, and then present a self-conducted empirical case study to verify the feasibility of our approach.

2. Hypertext resources

Grammis is a specialist hypertext resource that brings together terminological, lexicographical, and bibliographical information about German grammar. Initiated more than a decade ago, it combines traditional description of grammatical structures with the results of corpus-based studies and hypermedia design principles. Considering that the grammar of human languages is a highly complex scientific domain, the project authors use hypertext chunking and linking as well as multimedia extensions like spoken language excerpts and graphical explanations in order to address a broad target audience with heterogeneous foreknowledge. Their goal is to present a comprehensive overall picture of contemporary German grammar from a syntactic, semantic, and functional perspective. Today, Grammis[1] is the most prominent academic information system dedicated to German Grammar, with consistently more than 50,000 page impressions per month. ProGr@mm[2] is an e-learning system for schools, colleges, and uni-

[1] Short for: Grammatical Information System (*http://www.ids-mannheim.de/grammis/)*. The authors of this paper are members of the Grammis project team.

[2] Short for: Propaedeutic Grammar (*http://www.ids-mannheim.de/progr@mm/*)

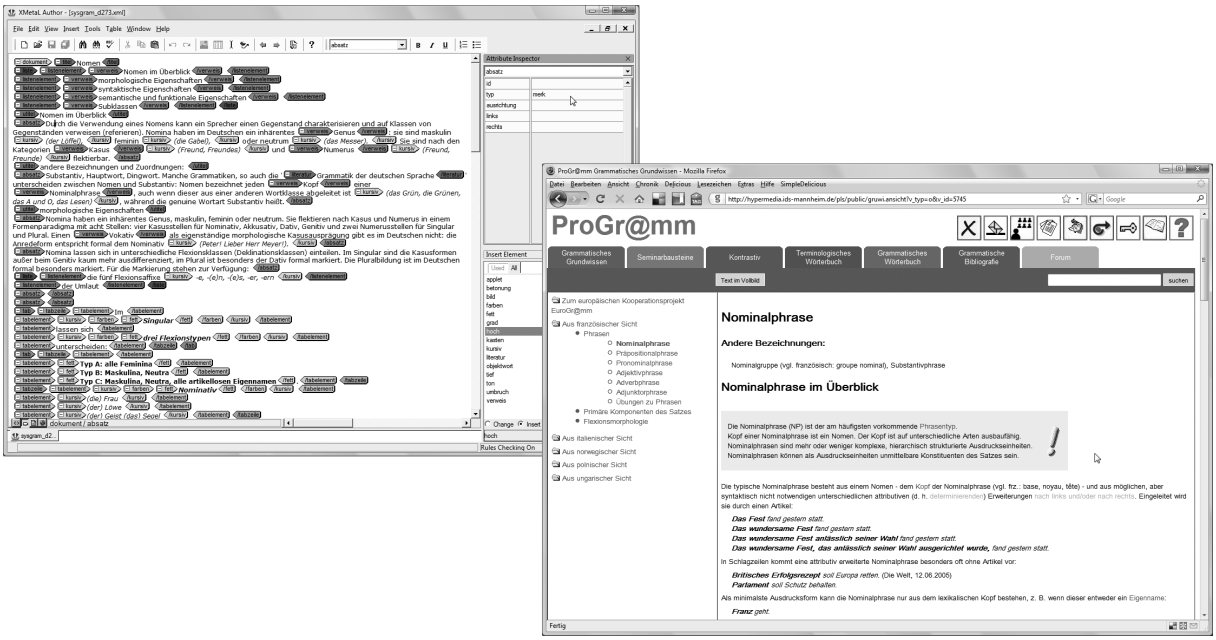


Fig. 1: XML stored inside the database (left) and converted to HTML (right)

versities, and didactically prepared for online learning. A special module covers selected grammatical topics from the perspective of other European languages and is well-suited for students and teachers of German as a foreign language. Functional add-ons are guided tours, personal notes, and discussion forums for the educational community.

From a technical point of view, both Grammis and ProGr@mm can be described as XML- and database-driven web information systems, whose semi-structured hypertext nodes (instances) conform to the Grammis Markup Language (GrammisML). GrammisML defines detailed constraints on the instance's logical structure, allowing for subsequent cross-media publishing ("one source fits all"). It provides conventional block elements like paragraphs, lists, or tables, as well as specific markup structures for the coding of grammatical metadata, typed hyperlinks, etc. Using a web-based authoring frontend, arbitrary keywords and object words/phrases for retrieval operations can be assigned manually. Parsing, analysis, and transformation of the hypertext resources are conducted using established technology like XPath, XQuery, XSQL, and XSL Transformation [11].

3. The domain ontology

Not just since the proclamation of the Semantic Web [2], semantic resources are among the most prominent add-ons and tools for information retrieval. Domain ontologies, organizing specialist terms (concepts) and their interconnections (relationships), can make a most valuable contribution to the analysis, classification, and finding of documents on the web — not least in the context of academic publications [3]. This is due to the both simple and unfortunate fact that scientific terminology is often far from being consistent. Especially in the field of linguistics, different theories, schools of thought, or even authors not only name things differently, but even assign varying meanings to identical terms. A semantically enriched retrieval application for the exploration of linguistic resources should incorporate these theory-related details so that it can offer appropriate solutions. As a consequence, we integrated a domain ontology for linguistic/grammatical terminology. The semiautomatic detection of concepts as well as the modeling of relationships has been conducted using statistical methods on large general language corpora and specialist language corpora.[3] Broadly speaking, in order to bring together theoretical desiderata with practical demands and limitations, we combine well-established standards of ontological engineering — e. g., the use of ISO-2788/ANSI Z39.19 compliant hyponymy/meronymy relationship types like Broader Term Generic (BTG) or Broader Term Partitive (BTP) — with terminological modeling principles — e. g., termsets, expanded by theory-related attributes and explicit linking of individual concepts belonging to different Termsets [1].

Figure 2 illustrates our ontology model. It covers three termsets, indicated by dotted border lines. The bottom termset contains the two concepts *Verbgruppe* and *Verbalphrase*, recognizable by rectangles with rounded corners. *Verbgruppe* is characterized by a theory-related attribute named *IDS'*, meaning that it is used primarily

[3] See [12] for a description of the ontology building process; [14] describes the ontology in greater detail.

when referring to the *IDS Grammar of German Language*. The concept *Verbalphrase* consists of four lexical entries:

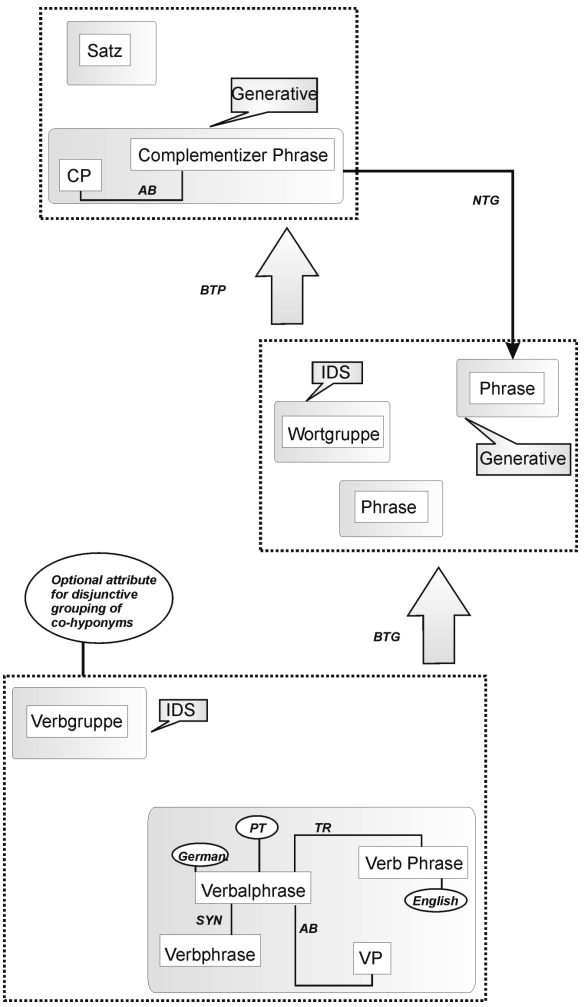


Fig. 2: Grammis ontology modeling structure

- *Verbalphrase* with a marker for Preferred Term (PT) and a language attribute (*German*)
- *Verbphrase* linked to the former by a synonymy relation (SYN)
- *VP* linked by a abbreviation relation (AB)
- *Verb Phrase* with language attribute (*English*) and translation relation (TR)

The termset is linked with its hyperonym by a Broader Term Generic (BTG) relation. In order to clarify the benefit of linking not only termsets, but also concepts, our example illustrates the relationships between *Phrase* (engl. *phrase*) and *Satz* (engl. *sentence*). Basically, the corresponding termsets are connected with the help of a Broader Term Partitive (BTP) relation (meronymy). Beyond this, since generative grammars usually classify sentences as phrases (*complementizer phrases*), only these two concepts — singled out by a theory-related attribute — are linked by a Narrower Term Generic (NTG) relation (hyponymy). This should facilitate communication between people or computers using different terminological systems.

4. The classification process

The goal of the classification process is to find terms (keywords) describing the content of a hypertext. We use the following information for our algorithm:

- The hypertexts contain XML-coded markup like paragraphs, lists, tables etc., but also specific grammatical metadata and links to the grammatical dictionary.
- For the classification process the hypertexts were lemmatized and part-of-speech tagged using the "TreeTagger" [10] and a training file for German.
- The source for possible keywords is our ontology, that can be accessed by functions such as „get hypernyms of a term *x* up to *n* levels" or „get synonyms of term *x*" etc.

4.1. The ranking algorithm

For the classification process we stored the hypertexts as a lemmatized word list which also contains the type of the paragraph the word is used in (title, subtitle or definition). We omitted words that are used in examples and tables: Examples contain object language that should not be used as a source for keyword candidates. Tables also often list object language and contain word chunks or fragments, because they are used for the presentation of inflection paradigms and the like. The basic idea of our classification algorithm is the following:

1. We select for each text all the terms that are also part of the ontology.
2. For each term, we assume broader terms one level above as additional keyword.
3. For each term, a rank is calculated that reflects its importance within the text. We use basically three factors to calculate importance:
   a. Frequency: More frequent terms are more important than rare ones.
   b. Position: If a term is mentioned in a title, subtitle or a definition or is used as a link to the grammatical dictionary, then it is supposed to be more important.
   c. Statistical significance: The relative frequency compared to the mean frequency of the term in all the other texts is calculated using a log-likelihood test.

These three factors are combined to an overall score. Frequency and position are calculated by counting the occurrences of the term in question multiplied by a weight depending on its position. In our standard procedure we used: titles = 6, subtitles = 4, definitions and „Merksätze" = 2, all other positions = 1. The statistical significance is calculated using the log-likelihood test [9]:

$$LL = 2*((a*\log (a/E_1)) + (b*\log (b/E_2)))$$

In this formula, a and b are the raw frequencies of the term in the text and the whole corpus respectively. E_1 and E_2 are the expected frequencies in the text and the whole corpus. The calculated value expresses the difference of the relative frequency to the total corpus. The higher the value, the higher is the significance of the term for the specific text. The following example demonstrates the difference between raw frequencies and relative frequency: Table 1 shows the frequencies and significance values of a hypertext node on valency ("Valenz").[4]

The keywords are ordered by their significance for the text (column „LLR"). Column „frequency" contains the raw frequencies, and „weighted frequency" stands for the frequencies weighted by the position in the text. The list also contains terms that are not mentioned literally, but are broader terms of a token („source term"). The most frequent term is *Valenzträger*, but according to the raw and the weighted frequency, *Valenzträger* would be on a lower rank. And vice versa: A very often used term like *Adjektiv* is not significant enough for a text on (verb) valency to rank in a top position ordered by significance.

[4] *http://hypermedia.ids-mannheim.de/pls/public/sysgram.ansicht?v_typ=d&v_id=2871*

4.2. The final ranking

The algorithm produces two different rankings: One ranking reflects the combination of frequency of the term and its position, the other ranking represents the significance of the term. Both aspects influence the final ranking. We combined the two rankings in the following way: The rank is transformed to a score by inverting the rank position. We then sum up the two scores and get a final ranking. In addition, we omit keywords with raw frequency 1 which tend to get very high LLR values but are not important enough to be included into the keyword list. When applying the algorithm to the valency hypertext node (see table 1 above for the raw frequencies), we get the final ranking as shown in table 2.

The number of keyword candidates depends on how congruent the two lists of the highest ranked terms are. Table 2 is based upon the combination of two top 10 lists and both lists contain more or less the same terms in different order. Therefore the merged list contains only two terms more than the two source lists. Intuitively, table 2 satisfyingly reflects the text about valency, similar to other hypertext nodes we evaluated manually. But, as described in the following section, we tried to further evaluate the lists for better results.

Table 1: Comparison of different measures for the frequency of terms in the valency hypertext node

ID	Type	keyword Candidate	Frequency	Weighted frequency	LLR	Source term
2871	d	Valenzträger	8	17	70,93	Valenzträger
2871	d	syntagmatische Beziehung	11	23	64,89	Valenz
2871	d	Valenz	11	23	64,89	Valenz
2871	d	Komplement	18	18	54,44	Komplement
2871	d	Leerstelle	3	3	26,21	Leerstelle
2871	d	Wortart	15	33	13,68	Verb
2871	d	Verb	15	18	13,68	Verb
2871	d	Nominalgruppe	2	2	13,33	Nominalgruppe
2871	d	Modifikator	10	14	10,94	Adjektiv
2871	d	Adjektiv	10	13	10,94	Adjektiv
2871	d	Satzadverbial	10	13	10,94	Adjektiv
2871	d	Nomen	10	13	9,85	Nomen
2871	d	Bedeutung	6	6	9,66	Bedeutung
2871	d	Verbvalenz	1	1	6,25	Verbvalenz
2871	d	Eigenschaft	4	4	4,58	Eigenschaft
2871	d	Prädikat	4	4	4,58	Eigenschaft
2871	d	Form	6	6	3,51	Form
2871	d	Ergänzung	1	1	3,34	Ergänzung
2871	d	Infinitivkonstruktion	1	1	3,15	Infinitivkonstruktion
2871	d	Anhebung	1	1	3,15	Infinitivkonstruktion
2871	d	Infinitkonstruktion	1	1	3,15	Infinitivkonstruktion
2871	d	Nominalphrase	4	4	2,65	Nominalphrase

Table 2: Final ranking of the terms in the text „Verbvalenz“

ID	Type	Keyword Candidate	Score
2871	d	Valenz	17
2871	d	syntagmatische Beziehung	17
2871	d	Wortart	15
2871	d	Valenzträger	14
2871	d	Komplement	12
2871	d	Verb	10
2871	d	Leerstelle	6
2871	d	Modifikator	5
2871	d	Nominalgruppe	3
2871	d	Satzadverbial	2
2871	d	Adjektiv	2

5. Evaluation of the classification results

5.1. Evaluation results

Some of the hypertext nodes are already classified by manually assigned keywords, using an uncontrolled vocabulary. These keywords are a measure to evaluate our automated classification and to experiment with different settings of the classification algorithm. Currently, the algorithm cannot cope with multi-word units, therefore we only analyze texts with one or more single-word keywords. Table 3 shows how the change of some parameters of the classification algorithm — e. g., weight of position (title, subtitle, etc.) — affects the matching of manually and automatically assigned keywords. We differentiate three matching levels: level 1 counts documents, that at minimum have one correspondence of manually and automatically assigned keywords. At level 2 at least 50 %, and at level 3 all of the manual keywords need to be matched by the automatic ones.

Table 3: Evaluation of the automatic assigned keywords

Version	Parameters	Level 1: Matching documents Min. 1 KW	Level 2: Min. 50 % KW	Level 3: Min. 100 % KW
1	*Default version* Weight of positions: titel = 10, subtitle = 4, definitions and „Merksätze“ = 2 Source lists: top 10	79,34 % 657/828	37,68 % 312/828	22,4 % 186/828
2	*More keywords* Equal to default version, but: Source lists: top 20	83,69 % 693/828	48,18 % 399/828	29,71 % 246/828
3	*More keywords* Equal to default version, but: Source lists: top 40	85,02 % 704/828	52,29 % 433/828	32,97 % 273/828
4	*More keywords* Equal to default version, but: Source lists: top 100	85,02 % 704/828	52,54 % 435/828	33,33 % 276/828
5	*Titles version* Equal to default version, but: titel = 30, subtitle = 10	79,59 % 659/828	38,53 % 319/828	23,19 % 192/828
6	*Versions with more keywords lead to the same results than versions 2–4 above*			
7	*No hypernyms* Equal to default version, but: Only literally used words are keyword candidates, no hypernyms.	79,10 % 655/828	37,07 % 307/828	22,46 % 186/828
8	*More hypernyms* Equal to default version, but: Hypernyms up to 2 levels above in the hierarchy	78,02 % 646/828	37,44 % 310/828	21,98 % 182/828
9	*More keywords* Equal to “more hypernyms” version (8), but: Source list: top 100	85,51 % 708/828	53,5 % 443/828	34,54 % 286/828

The evaluation illustrates two key issues for successful keyword detection:

- Getting all possible keyword candidates out of the text (tested with versions 1, 5, 7 and 8 in table 3).
- Putting the keyword candidates into the right ranking order, so that the top 10 ranking reflects the text content (tested with versions 2–4, 6 and 9 in table 3).

The first evaluation results are not too impressive: A 50 % matching of automatically and manually assigned keywords is only achieved at about 37 % of the documents (table 3, 1). About 80 % of the documents have at minimum one correspondence. Crucial seems the number of keywords that are included into the final list of keyword candidates. If this number is being increased, the matching scores also get better (table 3, 2–4). But even if the source lists contain 100 keyword candidates, only 52 % of the documents have matches at a 50 % level (85 % at level 1). If other parameters are changed, the score does not increase significantly: Neither accepting less nor more hypernyms (table 3, 7–8) has a substantial impact on the matching score. Only a higher weight of title positions (table 3, 5) slightly increases the score.

5.2. Discussion

These results interfere with our first impression when we intuitively evaluated documents without any manual keywords. Therefore, the manual classification process has to be examined. In 262 (32 %) of 828 documents, at minimum 80 % of all manually chosen keywords are not used at all within the hypertext nodes, even if the most narrow terms are taken into consideration. The reasons for that are manifold:

- Tagging issues influence the matching results: The TreeTagger does not lemmatize some plural forms (e. g., *Pronomina*) correctly. This leads to a mismatch in hypertext nodes where only the plural form is used.
- The fact that at the moment we cannot cope with multi-word units also affects the evaluation of the manual classification process.
- Our human classifiers tend to choose keywords that are neither mentioned in the hypertext node nor are close hypernyms of text words.

The above mentioned hypertext node ("Relativ-Elemente") also shows that different keynote annotators could disagree on the best solution (bad inter-rater reliability). *Pronomen* and *Wortart* are the manually assigned keywords, but another rater perhaps would also or instead set *Relativsatz*, *Relativ-Element* (as used in the title of the text) or something else as a keyword. Table 4 shows the automatically assigned keywords to the text.

Table 4: Final ranking of the terms in the text „Relativ-Elemente"

ID	Type	Keyword Candidate	Score
368	d	Phrase	11
368	d	grammatische Kategorie	10
368	d	Relativsatz	10
368	d	eingeleiteter Nebensatz	9
368	d	Einheitenkategorie	8
368	d	Relativ-Element	8
368	d	nicht-verbaler Ausdruck	7
368	d	Pronominalphrase	7
368	d	Einbettung	7
368	d	Proposition	6
368	d	Verkettungsverfahren	6
368	d	restriktiv	5
368	d	Präpositionalphrase	4
368	d	semantische Relation	4
368	d	phrasale Kategorie	2
368	d	Nominalphrase	1

6. Conclusion

The discussion shows the demand for a gold standard regarding the automatic detection of keywords for specialist texts. But the establishing of such standards seems difficult due to the fact that different (hypertext) publications even today mostly use different microstructures. An orientation to existing guidelines like TEI would possibly ease the determination of default settings for position parameters like title, subtitle, paragraph types, etc. Beyond that, controlled vocabularies for the manually assigned keywords — or, alternatively, the integration of user-independent data like social bookmark tags or folksonomies — would surely affect the congruity with machine-detected terms. Nevertheless, the first results of our ontology-based approach encourage for further application in the context of information retrieval and classification — and for methodological comparisons with other approaches for automatic keyword extraction.

References

1. *Beißwenger, M.* TermNet — ein terminologisches Wortnetz im Stile des Princeton WordNet // TU Dortmund: Institut für deutsche Sprache und Literatur, 2008.
2. *Berners-Lee, T. / Hendler, J. A. / Lassila, O.* The Semantic Web // Scientific American 284 (5), 2001. P. 34–43.
3. *Chiarcos, C.* An Ontology of Linguistic Annotations // LDV-Forum/Journal for Computational Linguistics and Language Technology, 2008. 23 (1). P. 1–6.
4. *Fensel, D.* Ontologies: Silver Bullet for Knowledge Management and Electronic Commerce // New York: Springer, 2001.
5. *Gottron, T. / Schneider, R.* A Hybrid Approach to Statistical and Semantical Analysis of Web Documents // Merabti, M. (Ed.): Proceedings of The Fifth IASTED European Conference on Internet and Multimedia Systems and Applications (EuroIMSA). Acta Press, 2009. P. 115–120.
6. *Gruber, T. R.* Ontology of Folksonomy: A Mash-up of Apples and Oranges // International Journal on Semantic Web & Information Systems, 2007. Vol. 3.2. P. 1–11.
7. *Gruber, T. R.* Ontology // Liu, L. / Tamer Özsu, M. (Eds.): Encyclopedia of Database Systems. New York: Springer, 2009.
8. *Maedche, A.* Ontology Learning for the Semantic Web // Kluwer Academic Publishers, 2002.
9. *Rayson, P. / Garside, R.* Comparing corpora using frequency profiling // Proceedings of the workshop on Comparing Corpora, 38th annual meeting of the Association for Computational Linguistics (ACL), Hong Kong, 2000. P. 1–6.
10. *Schmid, H.* Probabilistic Part-of-Speech Tagging Using Decision Trees // Proceedings of the International Conference on New Methods in Language Processing, Manchester, 1994. P. 44–49.
11. *Schneider, R.* E-VALBU: Advanced SQL/XML processing of dictionary data using an object-relational XML database // SDV — Sprache und Datenverarbeitung/International Journal for Language Data Processing, 2008. Vol. 32.1/2008. P. 33–44.
12. *Schneider, R.* A Database-driven Ontology for German Grammar // Rehm, G. / Witt, A. / Lemnitzer, L. (Eds.): Data Structures for Linguistic Resources and Applications. Proceedings of the Biennial GLDV Conference 2007. Narr, 2007. P. 305–314.
13. *Schneider, R.* Benutzeradaptive Systeme im Internet: Informieren und Lernen mit GRAMMIS und ProGr@mm // Mannheim: Institut für Deutsche Sprache (IDS), 2004.
14. *Sejane, I.* Database-Driven Access to Heterogeneous XML-Contents Using Domain Ontology of German Grammar // SDV — Sprache und Datenverarbeitung/International Journal for Language Data Processing, 2008. Vol. 32.1/2008. P. 71–87.
15. *Simperl, E.* Reusing ontologies on the Semantic Web: A feasibility study // Data & Knowledge Engineering, 2009. Vol. 68/10. P. 905–925.
16. *Staab, S. / Studer, R.* Handbook on Ontologies. International Handbooks on Information Systems // New York: Springer, 2009.
17. *Studer, R. / Davies, J. / Warren, P.* Semantic Web Technologies — Trends and Research in Ontology-Based Systems // John Wiley & Sons, 2006.
18. *Tran, T. / Cimiano, P. / Rudolph, S. / Studer, R.* Ontolgy-Based Interpretation of Keywords for Semantic Search // Proceedings of the 6th International Semantic Web Conference, 2007. P. 523–536.

Is a Companion a distinctive kind of relationship with a machine?

Yorick Wilks

University of Oxford

I start from the perspective of the EC COMPANIONS project, and set out its aim to model a new kind of human-computer relationship based on long-term interaction, with some tasks involved although the Companion is not inherently task-based, since there need be no stopping point to its conversation. Some demonstration of its functionality will be given but the main purpose here is an analysis of what it is people might want from such a relationship and what evidence we have for whatever we conclude. Is politeness important? Is an attempt at emotional sympathy important or achievable? Does a user want a consistent personality in a Companion or a variety of personalities? Should we be talking more in terms of a "cognitive prosthesis (or orthosis)?" — something to extract, organize, and locate the user's knowledge or personal information — rather than attitudes?

Introduction

The paper assumes that artificial Companions are on their way, and the interesting issues concern what they will be like. I am assuming two things here: first, that the robotic aspect is interesting but dispensable for this discussion. Dautenhahn has established interesting facts such as that people would prefer that robots approached them from the side rather than head on (Walters et al., 2009) and of course there will always be people who want things brought to them rather than getting up out of their chairs. But I will be concerned here with aspects of Companions such that embodiment is a secondary matter, provided they can converse with an owner and can reach out to the world via the internet for information and to establish action and control. Whether they are implemented as mobile phones, moving robots with prostheses, or just "warm furry handbags" with Wifi, is irrelevant to what I shall discuss here, though I shall often assume they can assume visual shape on a screen when necessary, but that is far short of a robot in any full sense.

Secondly, and still by way of scene setting, it is convenient to distinguish Companions from both (a) conversational internet agents that carry out specific tasks, such as the train and plane scheduling and ticket ordering speech dialogue applications back to the MIT ATIS systems (Zue et al., 1992), and also from (b) descendants of the early chatbots PARRY and ELIZA, the best of which compete annually in the Loebner competition (Loebner). These have essentially no memory or knowledge but are simple finite state response sets, although ELIZA had primitive "scripts" giving some context, and PARRY (Colby, 1971) had parameters like FEAR and ANGER that changed with the conversation and determined which reply was selected at a given point.

I take the distinguishing features of a Companion agent to be:

1) that it has no central or over-riding task and there is no point at which its conversation is complete or has to stop, although it may have some tasks it carries out in the course of conversation;
2) That it should be capable of a sustained discourse over a long-period, possibly ideally the whole life-time of its principal user;
3) It is essentially the Companion of a particular individual, its principal user, about whom it knows a great deal of personal knowledge, and whose interests it serves — it could, in principle, contain all the information associated with a whole life (in the sense of the Memories for Life consortium XXX);
4) It establishes some form of relationship with that user, if that is appropriate, which would have aspects associated with the term "emotion";
5) It is not essentially an internet agent or interface, but since it will have to have access

to the internet for information (including the whole-life information about its user) and to act in the world (as it is not a robot), we may as well assume its internet agent status, and so it should have, so far as possible, access to open internet knowledge sources.

By separating a Companion conceptually from both a task-based system and a chatbot, we immediately lose access to the two evaluation paradigms associated with those models of computer dialogue: the first in terms of task-completion (stickiness, timing, task success etc.) and the latter (usually) in terms of distinguishability from some set of human interlocutors. There is, at the moment, no clear evaluation paradigm for a Companion, even if we had one to evaluate, although there are ideas for creating one (Webb et al., 2010) and some of these have been applied to the first demonstrators from the COMPANIONS (Wilks, 2006) project itself.

Given this narrowing of focus in this paper, what questions then arise and what choices does that leave open? Here are some obvious questions that have arisen in the literature:

i) What aspects of a relationship should one aim at with a Companion, in terms of such conventional categories as emotion, politeness, affection etc.?
ii) Even if it is not a robot, in the sense of a free-moving entity, should it have a screen, and should it have a visible avatar for communication, whether human, animal or abstract?
iii) Does a Companion need a voice or could communication be by typing (such as on a mobile phone, laptop or PC)?
iv) Need it have one identifiable personality, or perhaps several, and should the user be able to choose the Companion's personality or shift between them if there are several? More generally, are the answers to these questions, and the settings and constraints they imply, dependent on the type of Companion — the domain or setting into which it is to be placed, or is there only one type of Companion subject to general constraints?
v) Does the Companion have any goals of its own, beyond carrying out a user's commands, if that is possible: should there be other overriding ethical constraints on what can be commanded, such as avoiding harm to the user, even if requested? Should there be ethical constrains *on the user* as to how the Companion can be treated?
vi) What safeguards are there for the information content of such a Companion, in the sense of controlling access to its contents for the state or a company, and how should a user best provide for its disposal in case of his/her own death or incapacity?
vii) What if anything does a Companion have to *know* to be plausible, and does it need a certain level of inference and memory capacity over the material of past conversations with the user?

Let us take these issues in turn.

1. Emotion, politeness and affection

Cheepen and Monaghan (1997) presented results some thirteen years ago that customers of some automata, such as ATMs, are repelled by excessive politeness and endless repetitions of "thank you for using our service", because they know they are dealing with a machine and such feigned sincerity is inappropriate. This suggests that politeness is very much a matter of judgment in certain situations, just as it is with humans, where inappropriate politeness is often encountered. Wallis (Wallis et al., 2001) has reported results that many find computer conversationalists "chippy" or "cocky" and suggests that this should be avoided as it breeds hostility on the part of users; he believes this is always a major risk in human-machine interactions.

We know, since the original work of Nass (Reeves and Nass, 1996) and colleagues that people will display some level of feeling for the simplest machines, even PCs in his original experiments, and Levy (2007) has argued persuasively that the trend seems to be towards high levels of "affectionate" relationships with machines in the next decades, as realistic hardware and sophisticated speech generation make machine interlocutors increasingly lifelike. However, much of this work is about human psychology, faced with entities known to be artificial, and does not bear directly on the issue of whether Companions should attempt to detect emotion in what they hear from us, or attempt to generate it in what they say back.

The AI area of "emotion and machines" is confused and contradictory: it has established itself as more than an eccentric minority taste, but as yet has nothing concrete to show beyond some better than random algorithms for detecting "sentiment" in incoming text (e. g. Wiebe et al., 2005), but even there its success is dependent on effective content extraction techniques. This work began as "content analysis" (Krippendorff, 2004) at the Harvard psychology department many years ago and, while prose texts may offer enough length to enable a measure of sentiment to be assessed, this is not always the case with short dialogue turns. That technology rested almost entirely on the supposed sentiment value of individual words, which ignores the fact that their value is content dependent. "Cancer" may be marked as negative word but the utterance "I have found a cure for cancer" is presumably positive and detecting the appropriate response to that rests on the ability to do information extraction beyond single terms. Failure to observe this has led to many of the classic foolishnesses

of chatbots such as congratulating people on the death of their relatives, and so on.

At deeper levels, there are conflicting theories of emotion for automata, not all of which are consistent and which apply only in limited ranges of discourse. So, for example, the classic theory that emotion is a response to the failure and success of the machine's plans (e. g. Marsella and Gratch, 2003) covers only those situations that are clearly plan driven and, as we noted, Companionship dialogue is not always closely related to plans and tasks. "Dimensional" theories (Cowie et al., 2001, following Wundt, 1913), display emotions along dimensions marked with opposed qualities (such as positive-negative) and normally distribute across the space emotion "primitives", such as FEAR, and these normally assigned by manual tagging, and they this rest, like the text-sentiment theories above, on pre-tagging and any learning based on them, of the sort that all learning engines perform over tag distributions (e. g. Ciravegna et al., 2004). The problem with this is that tagging for "COMPANY" or "TEMPERATURE" (in classic NLP) is a quite different task from tagging for "FEAR" and "ANGER". These latter terms are not, and probably cannot be, analyzed but rest on the commonsense intuitions of the tagger, which may vary very much from person to person — they have very low consilience between taggers.

All this makes many emotion theories look primitive in terms of developments in AI and NLP elsewhere. Appraisal Theory (Scherer et al, 2008) seeks to explain why individuals can have quite different emotional reactions to similar situations because they have appraised them differently, e. g. a death welcomed or regretted. Appraisal can also be of the performance of planned activities, in which case this theory approximates to the plan-based one mentioned above. The theory itself, like all such theories, has a large-commonsense component, and the issue for computational implementation is how, in assessing the emotional state of the Companion's user to make such concepts quantitatively evaluable. If the Companion conducts long conversations with a user about his or her life and, as in the case of the Senior Companion prototype (http://www.youtube.com/watch?v=-Xx5hgjD-Mw) which discusses photo images, then one might expect there to be ample opportunity to assess the user's appraisal of, say, a funeral or wedding by means of the application of the sentiment extraction techniques to what is said in the presence of the relevant image. In so far as a Companion can be said to have over-arching goals, such as keeping the user happy then, to that degree, it is not difficult to envisage methods (again based on estimates of the happiness, or otherwise, of the user's utterances) for self-appraisal by the Companion of its own performance and some consequent causal link to generated demonstrations of its own emotions of satisfaction or guilt.

Also relevant to what a Companion should be is the "Affective Loop" (AL) paradigm (Höök, 2004) which, like most of the theories of emotion discussed, and as John Wisdom once said of philosophical discoveries, are often the "running of a platitude up a flagpole": but AL is a useful corrective to some of the claims above and is intended essentially for computational implementation. It emphasizes:

- that there is a natural "feedback loop" involved in emotional interaction between parties and which is essential to any model
- but that emotional interaction and feedback should not be thought of as a matter of information transfer.
- it is much concerned with design, and the design of multimodal interactions of the display of color and sound — it is not essentially concerned with emotional language
- it emphases the relative vacuity of emotional labels or terms, as we did above, and peoples' intuitive understanding of them.

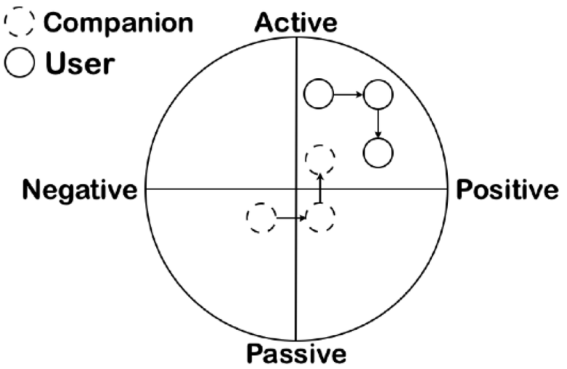
The notion of feedback is an old one going back to cybernetic ideas and in particular to Wiener's notion that activities like walking are only possible because of constant information feedback from the "servo" muscles in contact with the ground to the brain. Wiener was emphasizing information feedback, as opposed to the "haptic" transfer from muscles, but in a computational paradigm everything must at some stage bottom out in information. Speech act theory, too, arose from considerations of human interaction that were not based on conveying information in propositions, but rather "intentional" commitments, but those again have only been implementable in computers as forms of information.

Many of Höök's examples involve multi-modal devices such as smart phones where non-verbal signals are sent to create attitudes and feelings, or to signal those of the sender. The Nabaztag rabbit toy, originally used by the COMPANIONS project as an interface (http://www.nabaztag.com/en/index.html), in its original design glowed in a number of colors to indicate the feelings of the sender (e. g. blue for "sad") and two Nabaztags and their respective senders would be a paradigm AL in Höök's sense. There are many wholly conventionalised verbal feedback loops that cannot be divorced from emotion — certainly if a respondent fails to supply the correct response, from "How do you do" and "Good morning" in English to the potentially infinite "danke, bitte, danke, bitte…" cycle of giving thanks in German.

The importance of AL is that it makes emotion central, not peripheral, to communication and relationships and does not make language behavior central to emotional communication. Everyone knows that in relationships with pets, a central relationship for many people, this is the case: strong emotions are aroused, as well as consequent actions of e. g. stroking, but there is no verbal content. There have been a number of Japanese pet robot implementations, such as wriggly seal-like creatures with dozens of servo motors to give a life-like feel, and there is no doubt that a real form of human relationship is being modeled. Companions were always designed with the pet

analogy in mind, as in the phrase used early in the project of a Companion "being like a furry handbag", though language was always believed essential to the project.

In speaking of "language" and Companions, we have so far ignored speech, although that is a communication mode in which a great deal has been done to identify and, more recently, generate, emotion-bearing components (Luneski et al., 2008). Elements of the above approaches can be found in the work of Worgan and Moore (see e. g. Wilks et al., 2010), within the COMPANIONS project, where the is the same commitment to the centrality of emotion in the communication process, but in a form focusing on an integration of speech and language (rather than visual and design) technologies. The claim, not yet implemented, was conceived within the COMPANIONS project as a layer in a dialogue manager over and above local response management but one which would seek to navigate the whole conversation across a two-dimensional space onto which Companion and user are mapped using continuous values (rather than discrete values corresponding to primitive but unexplained emotional terms) but in such a way as to both respond to the a user's demonstrated emotion appropriately, but also — again, if appropriate or chosen by the user — to draw the user back to other more positive emotional areas of the two-dimensional space. It is not yet clear what the right mechanism should be for the integration of this "landscape" global emotion-based dialogue manager should be with the local dialogue management that generates responses and alters the world context: in the Senior Companion this last was sophisticated stack of networks (see Wilks et al., in press). In some sense, we are just looking for a modern and defensible interface to replace what PARRY had in simple form in 1971 when the sum of two emotion parameters determined which response to select from a stack of alternatives.



This last is a high level issue to be settled in a Companion's architecture and also, perhaps, to be under the control of the user, namely: should a Companion invariably try to cheer a user up if miserable — which is trying to "move" the user to the most naturally desirable (i. e. the top-right) quadrant of the space — or, rather, to track to the part of the space where the user is deemed to be and stay there in roughly the same emotional location — i. e. be sad with a sad user and happy with a happy one? There is no general answer to this question and, indeed, in an ideal Companion, which tracking method should be used would itself be a conversation topic e. g. "Do you want me to cheer you up or would you rather stay miserable?". In that sense, an AL is a platitude and everything depends on what kind of loop it is to be — itself a matter for negotiation.

2. What should a Companion look like?

I confess to an affection for a faceless Companion — the proverbial furry handbag, warm and light to carry, chatty but with full internet access and probably no screen. However, this may be a minority taste; after all, such a Companion could always take control of a nearby screen or a phone if it needed to show anything. If there is to be a face, the question of the "uncanny valley effect" always comes up, where it is argued that users are more uneasy the more something is very like ourselves (Mori, 1970). I personally do not feel this, indeed it cannot in principle apply to an avatar so good that one cannot be sure it is artificial, which is what I feel about the *Emily* from Manchester:

http://www.youtube.com/watch?v=UYgLFt5wfP4&feature=player_embedded#

http://www.surrealaward.com/avatar/3ddigital12.shtml

On the other hand, if the quality is not good, and in particular if the lip synch is not perfect, it may be better to go for an abstract avatar — the Companions logo was chosen with that in mind, and without a mouth at all. Non-human avatars seem to avoid some of the problems that arise with valleys and mixed feelings generally, and the best Companions demonstration video so far features Wigdog, a dog in a wig, who seems pretty popular:

http://www.youtube.com/watch?v=-Xx5hgjD-Mw

It may be worth making here a small clarification about the word "avatar" that sometimes distorts discussion in these areas: those working in computing the human-machine interface often use the word to mean any screen form, usually two-dimensional, that simulates a human being, but not any particular human being. On the other hand, in the virtual reality and game worlds, such as Second Life (http://secondlife.com/), an avatar is a manifestation of a particular human being, an alternative identity that may or may not be similar to the owner in age, sex, appearance etc. These are importantly different notions and confusion can arise when they are conflated or confused: in current COMPANIONS project demonstrations, for example, a number of avatars in the first sense are used to present the Companion's conversation on a computer

or mobile phone screen. However, in the case of a long-term computer Companion that could elicit, through prolonged reminiscence, details of its owner's life and perhaps train its own voice in imitation, since research shows that more successful computer conversationalists are as like their owners as possible. One might then approach the point where a Companion could approximate to the second sense of "avatar" above, namely an avatar of its owner, which it would progressively resemble, as dogs are said to do.

3. Voice or Typing to communicate with a Companion?

At the moment the limitation on the use of voice is twofold: first, although trained ASR for a single user — such as a Companion's user — is now very good and up in the high 90%, it still introduces uncertainty into understanding an utterance that is far greater than that of spelling errors. Secondly, it is currently not possible to store sufficient ASR software locally on a mobile phone to recognize a large vocabulary in real time; access to a remote server takes additional time and can be subject to fluctuations and delays. All of which suggests that typed input — though not TTS output — from a web-based Companion may have to use typed input in the immediate future, which is no problem for most mobile phone users who have come to find typed chat perfectly natural. However, this is almost certainly only a transitory delay as mobile RAM increases rapidly and the problem should not determine research decisions — there is no doubt that voice will move back to the centre of communication once storage and access size have grown by another order of magnitude.

4. One Companion personality or several?

Some (e. g. Pulman, in Wilks, 2010) have argued that having a consistent personality is a condition on Companionhood, but one could differ and argue that, although that is true of people — multiple personalities being a classic psychosis — there is no reason why we should expect this of a Companion. Perhaps a Companion should have a personality adapted to its particular relationship to a user at a given moment: Lowe (in Wilks, 2010) has pointed out that one might want a Companion to function as, say, a gym trainer, in which case a rather harsh attitude on the part of the Companion might well be the best one. If a Companion's emotional attitude were to (figuratively) move across a two dimensional emotion space (see diagram above) imitating or correcting what it perceived to be the user's state over time (as Worgan, see above, has proposed), then that shift in attitude might well seem to be the product of different personalities, as it sometimes can with humans.

It might be better, *pace* Pulman, to give a user access to, and some control over, the display of a multiple-personality Companion, something one could think of as an "agency" of Companions, rather than a single "agent", all of which shared access to the same knowledge of the world and of the state and history of the user.

5. Ethics and goals in the Companion

The last section is very close to the question of what goals a Companion can plausibly have, beyond something very general, such as "keep the user happy and do what they ask if you can", which are goals and constraints that directly relate to the standard discussions of the ethics a robot could be considered to have, a discussion started long ago by Asimov (1975). Clearly, there will be need for a Companion to have goals to carry out specific tasks: if it is to place a restaurant table booking on the phone for a user who has just said to it "Get me a table for two tonight at Branca around 8.30" — a phone request well within the bounds of the currently achievable technology — and the Companion will first have to find the restaurant's phone number before it phones and ask about availability before choosing a reservation time. This is the standard content of goal-driven behavior, with alternatives at every stage if unexpected replies are encountered (such as the restaurant being fully booked tonight). But one does not need to consider such goals as "goals of its own" since they are inferred from what it was told and are simply assumed, as an agent or slave of the user. But a Companion that finds its user not responding after some minutes of conversation might well have to take an independent decision to call a doctor urgently, based on a stored permanent goal about danger to a user who is unable to answer but is not asleep etc.

Asimov was concerned with the ethics of the robot and its doing no harm to its users, or indeed to anyone else — even if asked to do harm explicitly. These days one might also consider the point at which ill treatment of he Companion itself might be an ethical problem for the user: again, Nass' experiments revealing feeling or sympathy even for a criticized PC suggest these will not be too far in the future.

6. Safeguards for the information content of a Companion

Data protection, privacy, or whatever term one prefers, now captures a crucial concept in the new information society. A Companion that had learned intimate details of a user's life over months or years would certainly have contents needing protection, and many forces — commercial, security, governmental, research — might

well want access to it, or even to those of all the Companions in a given society. If societies move to a clear legal state where one's personal data is one's own, with the owner or originator having rights over sale and distribution of their data — which is not at all the case at the moment in most countries — then the issue of the personal data elicited by a Companion would automatically be covered.

If we ignore the issues of governments and national security — and a Companion would clearly be useful to the police when wanting to know as much as possible about a murder suspect, so that it might then be an issue of whether talking to one's Companion constituted any kind of self-incrimination, in countries where that form of communication is protected. Some might well want one's relationship to a Companion put on some basis like that of a relationship to a priest or doctor, or even to a spouse, who cannot always be forced to give evidence in common-law countries.

More realistically, a user might well want to protect parts of his or her Companion's information, or even an organized life-story based on that, from particular individuals: e. g. "this must never be told to my children, even when I am gone". It is not hard to imagine a Companion deciding whom to divulge certain things to, selecting between classes of offspring, relations, friends, colleagues etc. There will almost certainly need to be a new set of laws covering the ownership, inheritance and destruction of Companion-objects in the future.

7. What must a Companion know?

There is no clear answer to this question: dogs make excellent Companions and know nothing. More relevantly, Colby's PARRY program, the best conversationalist of its day (Colby, 1971) and possibly since, famously "knew' nothing: John McCarthy at Stanford dismissed PARRY's skills by saying :"It doesn't even know who the US President is", forgetting as he said it that most of world's population did not know that, at least at the time. On the other hand, it is hard to relate over a long term to an interlocutor who knows little or nothing and has no memory of what it or you have said in the past. It is hard to attribute personality to an entity with no memory and little or no knowledge.

Much of what a Companion knows that is personal it should elicit in conversation from its user; yet much could also be gained from publicly available sources, just as the current Senior Companion demo goes off to Facebook, independently of a conversation, to find out who its user's friends are. Current information extraction technology (e. g. Ciravegna et al., 2004) allows a reasonable job to be made of going to Wikipedia for general information when, say, a world city is mentioned; the Companion can then glean something about that city from Wikipedia and ask a relevant question such as "Did you see the Eiffel Tower when you were in Paris?" which again gives a plausible illusion of general knowledge.

John McCarthy always maintained that the real challenge for AI was not having exotic or detailed knowledge but common-sense knowledge, what exists below our levels of consciousness, such as that dropped thing fall, and fingers go into water when pushed but not into tables: all of what Hayes once called Naïve Physics (Hayes, 1978). Some of this can be coded in the inference rules a Companion will need, such as that sisters share parents, but much of it is below the level of straightforward rules, which is what led Dreyfus (1972) and others to argue that plausible AI would need the ability to learn as we do by growing up, rather than by existing forms of machine learning or hand-coding. However, the great improvements in such learning in recent years, from speech recognition to machine translation suggests that the jury is still out on this, even if the methods that have proved successful in computers are clearly not those humans themselves use.

8. A concrete Companion paradigm: the Victorian Companion

The subsections above are mini-discussions of some of the constraints on what it is to be a Companion, the subject of a recent book collection (Wilks, 2010). The upshot of those discussions is that there are many dimensions of choice, even within an agreed definition of what a Companion is to be, and they will depend on the user's tastes and needs above all. In the section that follows, I cut though the choices and make a semi-serious proposal for a model Companion, one based on a once well-known social stereotype.

In (O'Hara, in Wilks 2010) a colleague remarks that James Boswell was a clear case of the inaccurate Companion: his account of Johnson's life is engaging but probably exaggerated, yet none of that now matters. Johnson is now *Boswell's* Johnson, by and large, and his Companionship made Johnson a social property in a way he would never have been without his Companion and biographer. This observation brings out some of the complexity of Companionship, as opposed to a mere amanuensis or recording device, and its role between the merely personal and the social.

The first Artificial Companion is, of course, Frankenstein's monster in the 19C; that creature was dripping with emotions, and much concerned with its own social life:

Shall each man," cried he, "find a wife for his bosom, and each beast have his mate, and I be alone? I had feelings of affection, and they were requited by detestation and scorn. Man! you may hate; but beware! your hours will pass in dread and misery, and soon the bolt will fall which must ravish from you your happiness for ever (Shelley, 1831, Ch. 20).

This is clearly not quite the product that any modern COMPANIONS project is aiming at but, before just dismissing it as an "early failed experiment", we should take seriously the possibility, already touched on above, that things may turn out differently from what we expect and Companions, however effective, may be less loved and less loveable than we might wish. Newell has argued forcefully (e. g.in Wilks, 2010) that we must actually find out what kinds of relationship people want with Companion entities, as opposed to being technologists and just deciding a priori and then building what they believe people want.

It is no longer fashionable to explore a concept by reviewing its various senses, though it is not wholly useless either: when mentioning recently that one draft website for the COMPANIONS project had the black and pink aesthetic of a porn site, I was reminded by a colleague that the main Google-sponsored Companions site still announces "14.5 million girls await your call" and it was therefore perhaps not as inappropriate as I had first thought. Yet, for many, a Companion is still, primarily, a domestic animal, and it is interesting to note the key role pet-animals still play in the arguments on what it is, in principle, to be a Companion: especially the presence of the features of memory, recognition, attention and affection, found in dogs but rarely in snakes or newts.

I would also add that pets can play a key role in arguments about responsibility and liability, issues also raised already, and that dogs, at least under English common law, offer an example of an entity with a status between that of humans and mere wild animals: that is, *ferae naturae*, such as tigers, which the common law sees essentially as machines, and anyone who keeps one is absolutely liable for the results of its actions. Companions could well come to occupy such an intermediate moral and legal position (see Wilks & Ballim, 1990), and it would not be necessary, given the precedents with pets already available in law, to deem them either mere slaves or the possessors of rights like our own. Dogs are treated by English courts as potential possessors of "character", so that a dog can be of "known bad character", as opposed to a (better) dog acting "out of character". There is no reason to believe that these pet precedents will automatically transfer to issues concerning Companions, but it is important to note that some minimal legal framework of this sort is already in place.

More seriously, and in the spirit of a priori thoughts (and what else can we have at this technological stage of development?) about what a Companion should be, I would suggest we could profitably spend a few moments reminding ourselves of the role of the Victorian lady's Companion. Forms of this role still exist, as in a recent web posting:

Companion Job
posted: October 5, 2007, 01:11 AM
I Am a 47 year old lady looking seeking a position as Companion to the elderly, willing to work as per your requirements.I have been doing this work for the past 11 yrs.very reliable and respectful.
Location: New Jersey
Salary/Wage: Will discuss
Education: college
Status: Full-time
Shift: Days and Nights

But here the role has become more closely identified with caring and the social services than would have been the case in Victorian times, where the emphasis was on company, preferably educated company and diversion, rather than care. However, this was not always a particularly desirable or even tolerable role for a woman. Fanny Burney refers to someone's Companion as a "toad-eater" which Grose (1811) glosses as:

A poor female relation, and humble Companion, or reduced gentlewoman, in a great family, the standing butt, on whom all kinds of practical jokes are played off, and all ill humors' vented. This appellation is derived from a mountebank's servant, on whom all experiments used to be made in public by the doctor, his master; among which was the eating of toads, formerly supposed poisonous. Swallowing toads is here figuratively meant for swallowing or putting up with insults, as disagreeable to a person of feeling as toads to the stomach.

But one could nevertheless, and in no scientific manner, risk a listing of features of the ideal Victorian Companion:

1. Politeness
2. Discretion
3. Knowing their place
4. Dependence
5. Emotions firmly under control
6. Modesty
7. Wit
8. Cheerfulness
9. Well-informed
10. Diverting
11. Looks are irrelevant
12. Long-term relationship if possible
13. Trustworthy
14. Limited socialization between Companions permitted off-duty.

The Victorian virtue of Discretion here brings to mind the "confidant" concept that Boden (in Wilks, 2010) explicitly rejected as being a plausible one for automated Companions:

Most secrets are secret from some HBs [Human Beings] but not others. If two CCs [Computer Companions] were to share their HB-users' secrets with each other, how would they know which other CCs (i. e. potentially, users) to 'trust' in this way? The HB could of course say "This is not to be told to Tommy"... but usually we regard it as obvious that our confidant (sic) knows what should not be told

to Tommy — either to avoid upsetting Tommy, or to avoid upsetting the original HB. How is a CC to emulate that?

The HB could certainly say "Tell this to no-one" — where "no-one" includes other CCs. But would the HB always remember to do that?

How could a secret-sharing CC deal with family feuds? Some family websites have special functionalities to deal with this. E.g Robbie is never shown input posted by Billie. Could similar, or more subtle, functionalities be given to CCs?"

I think Boden brings up real difficulties in extending this notion to a computer Companion, but I do not think the problems are all where she thinks. I see no difficulty in programming the notion of explicit secrets for a Companion, or even things to be kept from specific individuals ("Never tell this to Tommy"). Companions will have less problems remembering to be discrete than people do, and I suspect there is less instinctual discretion that Boden suggests: people have to be told explicitly who to say what to in most cases, unless they are told to tell no one. In any case, much of this will be moot because Companions will normally deal only with one person — which is what makes their speech recognition problem so much easier, as we noted — they are trained for a single speaker — except when, say, making phone calls to an official, friend or restaurant, where they can try to keep the conversation to limited replies they can be sure to understand. The notion of a stored fact that must not be disclosed is simple to code, and the issue is wider in that the same fact must, to preserve the secret, not take part in inference processes either. If it is a secret that Tom is really a Russian, then the Companion should not do inferences like [IF X is of nationality Y THEN X will normally speak Y] and come out with an utterance like "I assumed Tom could speak Russian", which would rather give the game away via the reverse inference, in the hearer [IF X speaks Y THEN X may well be of nationality Y].

The interesting case Boden raises is that of Companions talking to each other, and this was presumably always a risk for Victorian ladies: that their human Companions would gossip behind their backs. For our Companions this seems a positive development that we might encourage: imagine the shy older person in a care home, too shy to approach another for a lunch together. This would be something best settled between their Companions, each knowing the tastes and habits of their owner, to whom the "date" could be presented as a fait accompli. Again, many Companion-to-Companion interactions will be between an individual's Companion and some form of "public Companion" such as one that takes restaurant bookings based on a user's tastes; or at a hospital where a hospital-Companion could triage incoming patients, who may not be articulate about their condition, on the basis of detailed knowledge of the user's medical records. When traveling, this Companion-to-Companion interaction in, say, a hospital could also combine with translation where the respective Companions worked out how to communicate across a language barrier.

In all these cases, Companion-to-Companion communication could be of obvious benefit to a user even if confidential information was at risk of disclosure: the user might have said "Never tell anyone I'm HIV positive" but in the hospital environment that constraint should obviously be overridden and the user's condition revealed. One could say at this point that secrets may be relative to a situation and that there may be nothing more complex in a Companion's guardianship of secrets than there is in explicit restrictions one could give to human hearers. The ultimate revelation of secrets by a Companion after a user's death is a wholly separate and complex subject. There are already on the market (e. g. Deathswitch: http://www.deathswitch.com/) products that save and reveal passwords and ultimate letters and secrets at death; this is undoubtedly an area with enormous possibilities as the Internet makes actual death less apparent and immediate in the electronic world than it is the real one (see also Wilks http://people.oii.ox.ac.uk/yorick/2007/01/24/death-and-the-internet/).

If the Victorian list of characteristics above is in any way plausible, it suggests an emphasis rather different from that current in much research on emotions and computers (e. g. the HUMAINE network at emotion-research.net) and their possible embodiments and deployments to a public. The emphasis in the list is on what the self-presentation and self-image of a possible, and tolerable, Companion should be; its suggestion is that overt emotion may not be what is wanted at all. I have never felt wholly comfortable with the standard Embodied Conversational Agent (ECA) approach in which if, an avatar "has" an emotion, it immediately expresses it, almost as if to prove the capacity of the screen graphics. This is exactly the sort of issue tackled by Darwin (1872) and such overtness can seem to indicate almost a lower evolutionary level than one might want to model, in that it is not a normal feature of much human interaction. The emotions of most of my preferred and frequent interlocutors, when revealed, are usually expressed in modulations of the voice and a very precise choice of words, but I realize this may be just cultural prejudice.

On the other hand, pressing the pet analogy might suggest that, if that is to be the paradigm, then overt demonstrations of emotion are desirable and sought by pet owners: dogs do not much disguise their emotions, and their positive emotions are often welcomed by owners. Language, however, does disguise emotion as much as it reveals it, and its ability to please, soothe and cause offence are tightly coupled with linguistic expertise — as opposed to the display of gestures and facial expressions — as we all know with non-native speakers of our languages who frequently offend, even though they have no desire to do so, and often have no awareness of the offence they cause. What name to call someone by, or whether or not to use vocatives like "Sir", "Mister", "Miss", "Missus" are enormously

complex matters, known intuitively to native speakers but not to outsiders, who are never taught them and have nowhere to go for advice or instruction. These are not cultural matters across space only, but also time: it was pointed out long ago that in the 19C male Cambridge undergraduates would walk arm-in-arm and call each other by their last names, without giving offence, whereas in the latter part of the 20C they would use first names — since last names would have given offence — and never be seen arm-in-arm!

I personally find the lady's Companion list above an attractive one: it eschews emotion beyond the linguistic, it implies care for the mental and emotional state of the user, and I would personally find it hard to abuse any computer with the characteristics listed above. It is no accident, of course, that this list fits rather well with the aims of the Senior Companion demonstrator in the COMPANIONS project already mentioned above. But the project first produced a Health and Fitness Companion (http://www.youtube.com/watch?v=KQSiigSEYhU&feature=related) for the more active, one sharing much of the architecture with the first, and one that would require something in addition to the list above: the "personal trainer" element of weaning, coaxing and threatening which adds something quite different to that list. and something very close to the economic-game bargain of the kind discussed in some detail by Lowe (in Wilks, 2010).

Many of the situations discussed above are, at the moment, wildly speculative: that of a Companion acting as its owner's agent, on the phone or World Wide Web, perhaps holding power of attorney in case of an owner's incapacity and, with the owner's advance permission, perhaps even being a source of conversational comfort for relatives after the owner's death. Companions may not all be nice or even friendly: Companions to stop us falling asleep while driving may tell us jokes but will probably shout at us and make us do stretching exercises. Long-voyage Companions in space will be indispensable cognitive prostheses (or, more correctly, orthoses) for running a huge vessel and experiments above any beyond any personal services — Hollywood already knows all that. All these situations are at present absurd, but perhaps we should be ready for them.

Acknowledgement

This work was funded by the Companions project www.companions-project.org) sponsored by the European Commission as part of the Information Society Technologies (IST) programme under EC grant number IST-FP6-034434.

References

1. *Colby, K. M.* "Artificial Paranoia." Artif. Intell. 2(1) (1971), pp. 1–2
2. *Cheepen, C. and Monaghan, J.* 1997, 'Designing Naturalness in Automated Dialogues — some problems and solutions'. In Proceedings 'First International Workshop on Human- Computer Conversation', Bellagio, Italy.
3. *Ciravegna, F., Chapman, S., Dingli, A. and Wilks, Y.* 2004. Learning to harvest the semantic web, in Proc. European Semantic Web Symposium (ESWS04)
4. *Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W. and Taylor, J. G.* 2001. Emotion recognition in human-computer interaction, Signal Processing Magazine, IEEE, 18(1), pp. 32–80.
5. *Höök, K.* (2004) User-Centred Design and Evaluation of Affective Interfaces, In From Brows to Trust: Evaluating Embodied Conversational Agents, Edited by Zsofia Ruttkay and Catherine Pelachaud, Kluwer's Human-Computer Interaction Series.
6. *Krippendorff, K. 2004. Content Analysis: An Introduction to Its Methodology. 2nd edition, Thousand Oaks, CA: Sage.*
7. *Levy, D.* 2007. Love and Sex with Robots: The Evolution of Human-Robot Relationships. London: Duckworth.
8. *Loebner:* http://www.loebner.net/Prizef/loebner-prize.html
9. *Luneski, A., Moore, R. K., & Bamidis, P. D.* (2008). Affective computing and collaborative networks: towards emotion-aware interaction. In L. M. Camarinha-Matos & W. Picard (Eds.), Pervasive Collaborative Networks (Vol. 283, pp. 315–322). Boston: Springer.
10. *Marsella, S. and Gratch, J.* (2003) Modeling Coping Behavior in Virtual Humans: Don't Worry, Be Happy. 2nd Int Conf on Autonomous Agents and Multiagent Systems (AAMAS), Melbourne, Australia, July 2003.
11. *Reeves, B., Nass, C.* 1996, The media equation: how people treat computers, television, and new media like real people and places, Cambridge: Cambridge University Press, 1996.
12. *Scherer, S., Schwenker, F. and Palm, G.* 2008. Emotion recognition from speech using multi-classifier systems and rbf-ensembles, in Speech, Audio, Image and Biomedical Signal Processing using Neural Networks, pp. 49–70, Springer: Berlin.
13. *Wallis, P., Mitchard, H., O'Dea, D., and Das, J.* 2001, Dialogue modelling for a conversational agent. In 'AI-2001: Advances in Artificial Intelligence', Stumptner, Corbett, and Brooks, (eds.), In Proceedings 14th Australian Joint Conference on Artificial Intelligence, Adelaide, Australia.
14. *Walters, M., Dautenhahn, K., te Boekhorst, R., Koay, K., Syrdal, D.*. 2009. An Empirical Framework for Human-Robot Proxemics.In Proc.AISB Convention 2009. www.aisb.org.uk/convention/aisb09/.
15. *Webb, N., Benyon, D., Hansen, P. and Mival, O.* (2010) Wizard of Oz Experiments for a Companion Dialogue System: Eliciting Companionable Conversation. Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC2010), Valletta, Malta. 2010.
16. *Wiebe, J., Wilson, T., and Cardie, C.* 2005. Annotating expressions of opinions and emotions in language. Language Resources and Evaluation, volume 39, issue 2–3, pp. 165–210.
17. *Wilks, Y.* 2006, 'Artificial Companions as a New Kind of Interface to the Future Internet. Oxford Internet Institute Research report No. 13 (Oxford Internet Institute). [Online], Available at: http://www.oii.ox.ac.uk/research/publications.cfm.
18. *Wilks, Y.* (ed.) (2010) Artificial Companions in Society: scientific, economic, psychological and philosophical perspectives. John Benjamins: Amsterdam.
19. *Wilks, Y., Catizone, R., Worgan, S., Dingli, A., Moore, R. K. and Cheng, W.* (in press) A prototype system for a conversational Companion for reminiscing about images. Computer Speech and Language.
20. *Wundt, W.* 1913. Grundriss der Psychologie, A. Kroner: Berlin.
21. *Zue, V., Glass, J., Goddeau, D., Goodine, D., Hirschman, L.* 1992. The MIT ATIS system, In Proc. Workshop on speech and natural language, Harriman, New York.

Abstracts

KEYWORD SEARCH USING SYLLABLE LATTICE

Aliyev R. M. (RomanAliyev@gmail.com), **Kheidorov I. E.** (igorhmm@mail.ru), **Yan Jinbing** (yjbemail@163.com),
Belarussian State University, Minsk, Belarus

Syllable lattice-based keyword search methods may help to overcome the problem of Out of Vocabulary (OOV) words and compensate the loss of search performance caused by recognition error. While there has been no effective search model in lattice-based search approaches, a syllable posterior probability-based search model is proposed. The model takes account of the lattice structure and syllable posterior probability. A search method based on the model is proposed. A series of experiments shows that our method is suitable for keyword search.

SEMANTIC SOURCES OF CONCESSION

Apresjan V. Ju. (valentina.apresjan@gmail.com), Russian Language Institute

The paper addresses the issue of concession as a complex derived meaning and analyzes its semantic origins. It also considers polysemy of concessive words and proposes semantic tools to distinguish among closely synonymous concessives derived from words with a non-concessive primary meaning. In particular, the following lexical items are analyzed: concessive conjunction "tol'ko" derived from a restrictive particle, and concessive conjunction/parenthetical word "pravda" derived from a factual noun. Their similarities and differences are analyzed in the light of the primary meanings of "tol'ko" and "pravda".

STATISTICAL DISTRIBUTIONS OF WORDS IN A COLLECTION OF RUSSIAN TEXTS

Baglei S. G. (baglei@galaktika.ru), **Antonov A. V.** (alexa@galaktika.ru), **Meshkov V. S.** (meshkov@galaktika.ru),
Sukhanov A. V. (sukhanov@galaktika.ru), Galaktika Corporation, Moscow, Russia

Statistical properties of texts have been widely studied in the fields of applied mathematics and linguistics. We explored statistical distribution of words in documents of a large collection of Russian texts using a probabilistic Bernoulli text generation process in our model. Unlike the traditional Bernoulli process, each document in the collection is considered as a finite text. We explored distributions of word frequencies in texts within a model representing a set of "bags-of-words". We plan to use the obtained results to elaborate a more realistic estimated probability of word generation in arbitrary Russian text with regard to word correspondence to the text collection.

SEMANTIC CORRELATES OF FORMAL VARIATION IN THE FIELD OF IDIOMATICS (THE OPERATION OF SUBSTITUTION)

Baranov A. N. (baranov_anatoly@hotmail.com), Institute of Russian Language, Moscow, Russia

The issue of formal variation of idioms is discussed. The paper focuses on the operation of substitution of different components on an idiom. A classification of different types of substitution operation is elaborated. It is hypothesized that formal variations of different kinds have specific semantic and discursive functions. Linguistic description of variation in the field of idiomatic presupposes an analysis of correlation between formal variation and meaning changes in an idiom. In is shown that substitution of the components of an idiom in most cases results in a generation of alternative semantic levels and, consequently, a linguistic play.

WWW STATISTICAL ESTIMATION OF THE FUNCTIONAL PROPERTIES OF LEXICAL ITEMS

Belikov V. I., Russian Language Institute n. a. V. V. Vinogradov, Russian Academy of Sciences
Akhmetova M. V., Journal "Zhivaia Starina" (Moscow, Russia)

The paper deals with the possibilities of using web-cite statistics for objective estimation of functional properties of vocabulary items: their stylistic status, territorial distribution, obsolescence of an item and its replacement by a new one, etc. The functional properties of particular words and phraseological units reveal themselves in their frequencies in different text arrays (classical vs. web-literature, official texts, weblogs, etc).

CREATING CONCEPTUAL GRAPHS AS ELEMENTS OF SEMANTIC TEXT LABELING

Bogatyrev M. Y.(okkambo@mail.ru), **Tuhtin V. V.**, Tula StateUniversity

Prospects of applying conceptual graphs as elements of semantic text labeling are considered. This kind of labeling is metadata that can be used to effectively solve some of the Text Mining problems. An algorithm for creating conceptual graphs is proposed and some results of its applications to modeling abstracts of scientific papers are presented.

A SPEECH CORPUS AS A TOOL FOR MONITORING AND FIXATION OF VARIOUS FORMS OF NATURAL LANGUAGE

Bogdanova N. V. (nvbogdanova_2005@mail.ru), **Asinovsky A. S.** (a.s.asinovsky@gmail.com),
Rusakova M. V. (mvrusakova@gmail.com), **Ryko A. I.** (aryko@yandex.ru),
Stepanova S. B. (stsvet_2002@mail.ru), **Sherstinova T. Yu.** (sherstinova@gmail.com),
Saint Petersburg State University, Russia

The paper concerns methodological principles and describes the technology of creation of the Corpus of Spontaneous Russian Speech and the structure of the database. Preliminary investigations based on Corpus material are briefly presented.

CROSSLEXICA: A LARGE ELECTRONIC DICTIONARY OF COLLOCATIONS AND SEMANTIC LINKS BETWEEN RUSSIAN WORDS

Bolshakov I. A. (bolshakov34@mail.ru), National Polytechnic Institute, Mexico

A large Russian electronic dictionary contains a vocabulary of 185,000 entries, 1.75 million collocations, 2 million semantic links, English translations of entry titles, and their morphoparadigms. It functions dialogically (for text editing or language learning) and is also accessible from external software for parsing, word sense disambiguation, detection & correction of malapropisms, steganography, etc.

CREATING A SEMANTIC DICTIONARY OF PREPOSITIONAL CONSTRUCTIONS ON THE BASIS OF THE UKRAINIAN NATIONAL LINGUISTIC CORPUS

Bugakov O. V. (ovbugakov@gmail.com), Ukrainian Lingua-Information Fund, NAS of Ukraine, Kiev, Ukraine

Search capabilities of the Ukrainian national linguistic corpus and linguistic databases built on its basis are examined. The structure of the semantic dictionary of prepositional constructions built in accordance with the theory of lexicographic systems is described. Key words: preposition, main word, dependent word, semantic state, electronic semantic dictionary of prepositional constructions.

MARKUP OF TEXT FRAGMENTS DURING CLASSIFICATION

Vasilyev V. G. (vvg_2000@mail.ru), Institute of Informatics Problems of the Russian Academy of Sciences (IPI RAN)

A comparative analysis of approaches to the selection of meaningful fragments of texts by using statistical methods of classification is presented. We consider new algorithms based on hidden Markov models covering the text by special hierarchical multiple fragments, as well as based on pre-segmenting the text into fragments without taking account of the information about the structure of classes.

RUSLED DICTIONARY AS TOOL FOR SEMANTIC STUDY

Voskresenskiy A. L. (avosj@yandex.ru), **Gulenko I. E.** (gig@yandex.ru), **Khakhalin G. K.** (gkhakhalin@yandex.ru)

The use of Russian sign language dictionary as an indicator of various Russian words meanings is described. This approach is enabled to more purposefully carry out analysis of context for word disambiguation.

THE DIGITAL RUSSIAN ASSOCIATIVE DICTIONARY OF SCHOOLCHILDREN

Goldin Valentin (goldinve@yandex.ru), **Martianov A. O.** (comrad-mao@mail.ru),
Sdobnova Alevtina (sdobnovaap@yandex.ru), Saratov State University

The paper deals with some ways of solving different issues of psycholinguistics, sociolinguistics and culturology based on the materials of the digital "Associative Dictionary of Schoolchildren of Saratov city and Saratov region".

ON THE NATURE OF SYNTACTIC POLISEMY

Grigorian E. L. (elena_grigorian@yahoo.co.uk), South Federal University, Rostov-on-Don, Russia

The analysis of variations of actant structures reveal the fact that most syntactic structures represent a set of semantic features which are not necessarily realized in every context. In many cases semantic distinctions are neutralized and the constructions differ only in communicative structure or style.

ON GESTURE–WORD CORRELATION (VOCAL GESTURE OH IN SPOKEN RUSSIAN)

Grishina Elena (rudi2007@yandex.ru), Institute of Russian Language, RAS, Moscow, Russia

The paper analyzes the usage of the vocal gesture Oh according to the data of the future Multimodal Russian Corpus (MURCO). The investigation is based on the analysis of the body and face movements that accompany this vocal gesture in the process of oral speech. As a result three meanings were detected 1) Oh as a deixis, 2) Oh as an interjection, and 3) Oh as a physiological exclamation.

UNIVERSAL DICTIONARY OF CONCEPTS

Dikonov V. G. (dikonov@iitp.ru), **Boguslavsky Igor M.** (bogus@iitp.ru),
Institute for Information Transmission Problems, Russian Academy of Sciences

A universal dictionary of concepts, developed as a part of the ongoing effort to create a semantic intermediary language for global information exchange, is presented. The article describes basic principles and contents of the dictionary and outlines the current state of the project. The dictionary can evolve into an open and freely available language-neutral resource with many potential applications. For example, the extensible dictionary of concepts can serve as a pivot to uniformly record and link meanings of words of different languages and facilitate creation of bi- and multilingual dictionaries. Another possible use is word sense markup of corpora. The dictionary of concepts is going to be linked at the word sense level with lexicons of major world languages including Russian, English, Spanish, French, Arabic, Hindi, etc.

RUSSIAN NET, GERMAN NEIN , ENGLISH NO: CONTRASTIVE SEMANTIC ANALYSIS WITH PARALLEL CORPORA

Dobrovol'skij D.O. (dm-dbrv@yandex.ru), **Levontina I. B.** (irina.levontina@mail.ru),
Russian Language Institute, Russian Academy of Sciences

'No' seems to be a very simple and universal idea. However, surprisingly enough, the German word nein or the English no are not always good equivalents for the Russian word net, and vice versa. Parallel corpora show that in many cases net is translated differently, even though the respective phrase with nein/no is acceptable. And we often see net in Russian translation instead of some other units. We assume that such lack of coincidence must have certain semantic reasons. They are probably rooted in semantic differences between net and nein/no. In our paper we try to reveal these reasons.

NATURAL LANGUAGE QUERY PROCESSING FOR SEARCH ENGINE BASED ON LINGUISTIC ANALYSIS

Ermakov A. E. (ermakov@metric.ru), **Pleshko V. V.** (vp@rco.ru), RCO Ltd (www.rco.ru), Moscow

A new method of transforming natural language queries into search engine language queries is described, which is based on the automatic analysis of syntactic relations between words and their representation as relevant search engine language operators saving the meaning of an original query to the extent possible.

ON THE NOTION OF SEMANTIC SHIFT

Zalizniak Anna A. (anna-zalizniak@mtu-net.ru), Institute of linguistics, Russian Academy of Sciences

The paper deals with the notion of "semantic shift" as a category of semantic typology and the unit of the "Catalogue of semantic shifts in the languages of the world"; it reflects some results of the work on a project, realized in the Institute of Linguistics, Russian Academy of Sciences, by a group of linguists (Anna A. Zalizniak, Maria Bulakh, Dmitriy Ganenkov, Ilya Gruntov, Timur Maisak and Maxim Russo). The problem of identification of semantic shifts in cases of syncretism (semantic generality) is discussed in more detail.

<STRIKE>I'VE NEVER TOLD THAT</STRIKE>: ABOUT LITURATIVES, STRIKEOUTS OR IMAGINARY TEXTS

Zanegina N. N. (zanegina@list.ru), The Institute of the Russian Language

This paper deals with linguistic peculiarities of strikeout texts — their semantics and syntax. These texts are very often used in Internet communication.

AN APPROACH TO AUTOMATED ONTOLOGY BUILDING IN TEXT ANALYSIS PROBLEMS

Zakharova I. V. (iren@csu.ru), **Gorodechnyj P. P.** (petr.gorodechnyj@edu.csu.ru),
Chelyabinsk State University, Dept. of mathematics

An approach to how to automatically build an ontology for complex tasks of full-text document classification using UDC is discussed.

UNIVERSAL SYNTAX ANNOTATION SYSTEM OBJECTATE

Zobnin A. I. (Alexey.Zobnin@gmail.com), Lomonosov State University,
Russian Language Institute n.a. V. V. Vinogradov, Russian Academy of Sciences
Sakharova A. V. (nenen@mail.ru), Russian Language Institute n.a. V. V. Vinogradov, Russian Academy of Sciences

The object model and the features of the Universal text annotation system ObjectATE are described. This system is used in Vinogradov Institute of Russian language of RAS for semimanual morphological and syntactical annotation of ancient manuscripts. It allows the user to define his own annotation models by describing classes, add-ins, fields and relations in the metadata layer (for example, for syntax markup).

EVERYDAY TERMINOLOGY. IN PURSUIT OF STANDARDS

Iomdin B. L. (iomdin@ruslang.ru), Russian Language Institute n.a. V. V. Vinogradov, Russian Academy of Sciences

The paper is devoted to the vocabulary describing everyday life artifacts. This vocabulary is shown to be treated very differently in dictionaries, production standards, and usage; henceforce, unified lexicographic definitions of words belonging to this vocabulary are hardly possible at all. A draft project of an explanatory and encyclopedic thesaurus of everyday life terminology is presented.

SYNTACTIC CORRELATES OF PROSODICALLY MARKED ELEMENTS OF THE SENTENCE AND THEIR ROLE IN THE TASKS OF TEXT-TO-SPEECH SYNTHESIS

Iomdin L. L. (iomdin@iitp.ru), Institute for Information Transmission Problems, Russian Academy of Sciences
Lobanov B. M. (lobanov@newman.bas-net.by), United Institute of Informatics Problems,
National Academy of Science of the Republic of Belarus

The paper describes a feasibility study of using syntactic parsing of written text at an initial stage of text-to-speech synthesis algorithm. An attempt has been made to establish correlations between the elements of an automatically created dependency tree structure of a sentence, on the one hand, and prosodically strong elements of this sentence, on the other hand. First experimental results show that the approach may be effective.

PROSODIC TRANSCRIPTION: LEVELS OF DETAIL

Kibrik A. A. (kibrik@comtv.ru), Institute of Linguistics RAS, **Khudyakova M. V.** (mariya.kh@gmail.com), Lomonosov Moscow State University, **Kodzasov S. V.** (sankod@yandex.ru), Lomonosov Moscow State University

In the book Kibrik and Podlesskaya (eds.) 2009, a prosodically oriented system of discourse transcription for spoken Russian was proposed. In this paper a number of extensions for that system are suggested, such as the distinction between expiratory and pitch accents, a more detailed account of pitch accents, interval of tone in an accent, dynamic vowel doubling, etc.

TOWARDS THE PROBLEM OF LINGUISTIC VARIABILITY: A MULTIFACTOR SECOND-ORDER CALCULUS METHOD

Kibrik A. E., Lomonosov Moscow State University, Russia

Clausal coordination is studied in 23 related Daghestanian idioms. Clausal coordination is extremely variable across this language sample: there is not a pair of idioms with identical coordinate clausal constructions. At first sight, the choice of formal coding technique used by specific idioms appears random and chaotic. Such situation creates irresolvable theoretical difficulties. Neither the traditional method of classification nor the structural calculus method are helpful.
In the paper an alternative method is employed. It can be called the multifactor second-order calculus method. A calculus of coordinate constructions is implemented at the level of parameterized principles and strategies predetermining specific coordinate constructions, rather than at the level of coordinate constructions themselves.

THE METHOD OF AUTOMATED SYNTAX SEGMENTATION RULES GENERATION

Klyshinsky E. S. (klyshinsky@mail.ru), Keldysh Institute of Applied Mathematics RAS
Manushkin E. S. (EugeneLebowsky@mail.ru), Moscow State Institute of Electronics and Mathematics

The paper proposes a method of automated generation of syntax segmentation rules. The method is based on FIRST, LAST, FIRST2 and LAST2 sets calculated for existing BNF grammars describing the rules for syntax analysis of natural languages texts.

SYNTACTIC INCOMPATIBILITY AS A PROPERTY OF THE LINEAR ORGANIZATION OF A RUSSIAN SENTENCE

Kobzareva T. Yu. (stamstam@mtu-net.ru), Russian State University for Humanities, Russia

The paper considers a property of the linear organization of sentence in Russian, the so-called syntactic incompatibility, or impossibility of simultaneous appearance of some components in its fragments set by punctuation marks or coordinative conjunctions. The property can be taken into account at different stages of automatic analysis.

SEMANTICS OF THE VERB PONIMAT': FROM PROPOSITIONAL TOWARDS INTERPERSONAL ATTITUDE

Kobozeva I. M. (kobozeva@list.ru), Lomonosov Moscow State University, Moscow, Russia

The Russian verb ponimat' 'understand' in constructions with a personal direct object is studied. 6 of its readings, corresponding to different intentional states (rational, emotional, interpersonal) are explicitly defined. The emergence of non-rational readings is explained on the cognitive basis.

THE DATABASE ON INTONATION OF RUSSIAN NARRATIVE TEXTS

Kodzasov S. V. (sankod@philol.msu.ru), **Arkhipov A. V.** (arxipov@philol.msu.ru), **Zakharov L. M.** (leon@philol.msu.ru),
Krivnova O. F. (okri@philol.msu.ru), Lomonosov Moscow State University, Moscow, Russia

The paper represents the results obtained at the 2nd stage of development of the DB "Intonation of the Russian informative and narrative texts". This stage opened the 2nd triennial cycle of inquiry into Russian intonation.

DETECTION OF NOMINALIZED STRUCTURES IN PARALLEL PATENT TEXTS IN RUSSIAN AND IN GERMAN

Kozhunova O. S. (kozhunovka@mail.ru), Institute for Informatics Problems of the Russian Academy of Sciences

In the paper nominalization in bilingual situation (Russian-German) involving comparative study results for three languages (Russian, English, and German), approach of parallel texts identification for patent sphere and transformation types have been analyzed.

PARENTHESES IN RUSSIAN IDIOMS

Kozerenko A. D. (akozerenko@mail.ru), V. V. Vinogradov Russian Language Institute, Russian Academy of Sciences

The paper offers a semantic analysis of Russian idioms containing the word parentheses. A paradoxical fact is observed that two idioms that in Russian sound as enclose in parentheses and put outside the parentheses have the same meaning. The clue is given and other Russian idioms containing the word parentheses are examined.

PAUSES AFTER POSTPOSITIONS AND TOPICAL PARTICLE WA IN JAPANESE: A CORPUS STUDY

Komarova A. D. (komarovochka@gmail.com), Russian State University for Humanities, Russia

This research discusses pauses after postpositions and topical particle wa and before them in Japanese. It aims to find out how frequent and probable these pauses are, their usual length and if they differ depending on the syntactic position.

A CORPUS STUDY OF PAUSATION AT SYNTACTIC BOUNDARIES: WHY PAUSES DO NOT ALWAYS APPEAR WHERE WE EXPECT THEM

Korotaev N. A. (n_korotaev@hotmail.com), Russian State University for Humanities (RSUH)

The so-called ideal delivery presupposes a pause at every elementary discourse unit boundary. Natural discourse, however, provides numerous examples when such pauses are missing. The paper reports a corpus-based study of these cases in spoken Russian. It is argued that they are characterized by a high degree of semantic integration, which correlates with syntactic and prosodic properties of the examined sequences. For instance, absence of pauses is intrinsic to most complex clauses. Analyzing a corpus of night dream stories, it has also been found that the ratio between boundaries with and without pauses varies appreciably from one story to another.

PATTERNS OF EMOTIONAL REACTIONS IN COMMUNICATION: PROBLEMS OF CORPORA STUDIES AND APPLICATION TO COMPUTER AGENTS

Kotov A. A. (kotov@harpia.ru), Russian State University for the Humanities

We study cognitive architecture of computer agents, simulating emotional speech behavior, and changing their mood in time. Basing on a multimodal corpus (records of university exams) we study sequences of contrastive emotional reactions and the possibility to apply the sequences to computer agents.

CITIZEN-INSTITUTION NON-MEDIATED DIALOGUE: THE RUSSIAN DIRECT LINE CASE

Cotta Ramusino P. (paola.cottaramusino@unimi.it), University of Milan, Italy

This paper analyses a specific kind of institutional discourse: Russian Direct Line. It aims to give account of interactional strategies used by subordinate participants of the given interaction. It tries to investigate how "naïf" interviewers, who are not familiar with strategies regulating a neutral or "neutralistic" position, manage to avoid possible consequences of their own speech acts, by using pragmatic and metapragmatic acts, basically aimed at downgrading.

THE NONVERBAL BEHAVIOR OF PEOPLE OF DIFFERENT CULTURES IN A DIALOG I: FINNISH AND RUSSIAN GESTURE SYSTEMS

Kreydlin G. E. (gekr@iitp.ru), Russian State University for Humanities

The paper presents some reflections of the so-called exterior observer about Finnish nonverbal semiotic culture, some nonverbal signs and models of Finnish dialog behavior. Corresponding Russian nonverbal data are given for comparison.

ON THE SEMANTIC CLASSIFICATION PROGRAM ProSeCa: THEORETICAL AND PRACTICAL ASPECTS

Kretov A. A. (a_a_kretov@rambler.ru), Voronezh State University, Russia
Rafaeva A. V. (anna_raf@rambler.ru), Lomonosov Moscow State University, Russia

A modified version of E. Kuznetsova's definition-based semantic identification method is proposed. The main point of it is that lexical semantics is concentrated in the most common nouns. A computer program of semantic classification is described. Perspectives of using and developing the program are outlined.

"QUASI-CORPUS" INVESTIGATION OF LEXICAL PRODUCTIVITY OF NON-TRIVIAL BASIC DIATHESES OF RUSSIAN WITH SPECIAL REGARD TO S. I. OZHEGOV'S DICTIONARY OF RUSSIAN

Krylov Sergey A. (krylov-58@mail.ru), Institute of Oriental Studies of Russian Academy of Sciences, Moscow;
Institute of System Analysis of Russian Academy of Sciences, Moscow

"Quasi-corpus" linguistics allows the investigation of both primary and secondary information sources (like grammars and dictionaries). The paper studies the statistics of grammatical data (on government patterns, transitivity, impersonality etc.) in the text of S. I. Ozhegov's "Dictionary of Russian" (1989).

THE ADJECTIVES WITH MEANING OF HIGH AND LOW TEMPERATURE AND LINGUISTIC ESTIMATION OF TEMPERATURE

Krylova T. V., V. V. Vinogradov Russian Language Institute, Russian Academy of Sciences

In this article the adjectives холодный, прохладный, горячий, жаркий, теплый are considered. In the first part we analyze their division to groups. In the second part we consider their combinations with adverbs of degree. We advance the hypothesis that many differences in using of temperature adjectives are caused by difference in linguistic estimation of high and low temperature. In conclusion the same idea is illustrated by the material of verb with meaning of temperature.

SEMIAUTOMATIC MARKING TOOLS FOR THE RUSSIAN MULTIMEDIA CORPUS (MURCO)

Kudinov M. S. (peshka1@mail.ru), MSU, **Grishina E. A.** (rudi2007@yandex.ru), Institute of Russian Language, Moscow

The paper describes two workbenches for corpus markers: a speech act marker's workbench (Marker) and a gesture marker's workbench (GesturesMarker). These programs allow the annotator to describe in quick and uniform manner Russian gesticulation and speech acts used in Russian spoken language.

MEANS FOR TUNING OF THE "SEMANTIX" LINGUISTIC PROCESSOR TO SUBJECT FIELDS

Kuznetsov I. P. (igor-kuz@mtu-net.ru), Institute of Informatics problems, Moscow, Russia
Efimov D. A. (d.efimov@synsys.ru), ZAO Synergetic Systems, Moscow, Russia

The linguistic processor "Semantix" for automatic formalization of natural language texts is presented It extracts data on user objects, their links and actions from texts. The processor uses special tools and methods for tuning to new subject fields. As an example the process of tuning for the text corpus about monuments is considered.

THE SEMANTIC DATABASE OF VERBAL ADJECTIVES: STRUCTURE AND TYPES OF INFORMATION

Kustova G. I. (galinak03@gmail.com), Moscow State Pedagogical University

The paper discusses the issues of elaboration of an electronic semantic dictionary (database) of Russian verbal adjectives (like vkhodnoj, lechebnyj, osvetitel'nyj etc). The topics considered include: a) the correlation between the verbal adjective and the verbal situation and the possibilities of expressing verbal arguments, e.g. stiral'naja mashina ('washing machine', instrument), vs. stiral'nyj poroshok ('washing powder', means); b) the correlation between the semantic class and the functional predicate of a noun and the semantic model of combinations like «verbal adjective + noun»; c) information types in the database; d) specification of semantic marking in the dictionary of the National Corpus of Russian language.

THE APPROACH TO CREATION OF MULTILINGUAL PARALLEL CORPORA OF WEB PUBLICATIONS

Lande D. V. (dwl@visti.net), **Zhigalo V. V.** (vladlen@visti.net), ElVist iInformation centre, Kiev, Ukraine

An algorithm of creating bilingual parallel corpora of documents from web publications is described. The algorithm uses frequency morphological dictionaries and empirical statistical properties of texts. An approach of homonymy resolution by means of statistical approach is presented, which allows choosing the most frequent normal forms. The algorithm has been developed as a software complex and integrated into the InfoStream system of content monitoring. As a result of algorithm operation aimed to determine basic word forms, a bilingual parallel corpus of electronic texts from web publications that contains more than 450 000 pairs of documents.

AN EDITOR OF AUGMENTED TRANSITION NETWORKS WITH A GRAPHICAL USER INTERFACE

Lebedev A. S. (andremoniy@gmail.com), Moscow State Institute of Electronics and Mathematics

The problem of semantic search is considered on an example of search for abstracts. An approach to the creation of a linguistic processor using augmented transition networks, inserted graphs, and arrangement of objects based on their descriptive part is proposed.

THE PROBLEM OF THE «Ë»-HOMOGRAPHS RESOLUTION IN TEXT-TO-SPEECH SYNTHESIS

Lobanov B. M. (lobanov@newman.bas-net.by),
United Institute of Informatics Problems, National Academy of Science of the Republic of Belarus

The problem of adequate ambiguity resolution in text-to-speech synthesis, for a special case of graphic homonymy related to the letter Ë is considered. Statistical characteristics of homographic pairs including Ë homographs and distributions among the frequent pairs of such homographs are investigated. The methods of resolution for the highly frequent homographic pair «ВСЁ» and «ВСЕ» are discussed.

SUMMARIZATION OF NEWS CLUSTERS BASED ON THEMATIC REPRESENTATION

Loukachevitch N. V. (louk@mail.cir.ru), **Dobrov B. V.** (dobroff@mail.cir.ru), Research Computer Center of M. V. Lomonosov Moscow State University (MSU NIVC); NCO Center for Information Research

The paper describes a technology of multi-document summarization, based on news cluster topical structure, lexical cohesion modelling and thesaurus descriptions of lexical senses. Lexical knowledge helps to improve cohesion and recall of a summary and reduce repetitions.

RUSSIAN FRAMENET: TOWARDS A CORPUS-BASED DICTIONARY OF CONSTRUCTIONS

Lashevskaya O. (olesar@mail.ru), **Kuznetsova Ju.** (julia.kuznetsova@uit.no), University of Tromsø (Norway)

The paper presents our basic approach to creating a FrameNet-oriented resource for Russian language, which involves extracting sampling from the Russian National Corpus and adding a layer of semantic and syntactic annotation. We discuss aims and methods of the project and give several examples of argument labeling in the dictionary and in the companion corpus.

NAMES OF BODY PARTS FROM THE VIEWPOINT OF TOPOLOGY

Makhova A. A. (discourse@yandex.ru), **Lyashevskaya O. N.** (olesar@mail.ru), **Desyatova A. V.** (patine@gmail.com)

The paper describes Russian names of body parts through the notion of topological type as introduced by L. Talmy. The corpus analysis of collocation with adjectives of shape and dimension makes it possible to define a number of topological types of body parts, such as juts, rods etc. and identify some peculiarities of their spatial perception.

AUTOMATIC ANALYSIS OF TERMINOLOGY IN THE RUSSIAN TEXT CORPUS ON CORPUS LINGUISTICS

Mitrofanova O. A. (alkonost-om@yandex.ru),Saint Petersburg State University, Russia
Zakharov V. P. (vz1311@yandex.ru), Saint Petersburg State University; Institute of Linguistic Studies, Russian Academy of Science

The paper presents the results of semi-automatic analysis of terminology in the Russian text corpus on Corpus Linguistics. Special attention is given to extraction of one-word and multi-word terms as well as to the use of lexical-grammatical patterns in the description of term structure and contexts of use.

AN EXPERIENCE OF CREATION OF THE NATIONAL CORPUS OF DAGESTAN LANGUAGES

Mutalov R. O. (mutalovr@mail.ru), Dagestan State University, Makhachkala

Problems and prospects of national corpora of six literary languages of Dagestan, created in the Dagestan State University, are considered. Special attention is given to the creation of a system of automatic markup of texts and digitization of printed texts.

COREFERENCE ANNOTATION IN PRAGUE DEPENDENCY TREEBANK

Nedoluzhko A. (nedoluzko@ufal.mff.cuni.cz), Charles University, Prague, Czech Republic

The paper presents the pattern for annotating coreferential relations on the PTD corpus. Three levels of annotation are discussed: annotating grammatical coreference (the antecedent is calculated according to the grammar rules of a given language); annotating textual pronominal coreference; an extended pattern for annotating nominal textual coreference and associative anaphora. The first two (grammatical coreference and pronominal coreference) have been annotated on the whole PDT corpus, whereas the nominal coreference and assosiative anaphora are currently in the focus of the author's research. Certain complicated cases are going to be discussed and first results of the research presented.

SEGMENTATION OF ORAL NARRATIVE DISCOURSE AND ILLUSTRATIVE GESTURES: VISUAL CLUES AS SEGMENT MARKERS

Nikolaeva Y. V. (lis_julia@list.ru), Lomonosov Moscow State University, Russia

The paper is devoted to the interrelations between speech accompanying gestures and the discourse structure. The main aim was to find out how different characteristics of illustrative gestures mark discourse segment boundaries.

MODELS AND METHODS OF PUNCTUATION USE IN RUSSIAN LANGUAGE SYNTAX PARSING

Okatiev V. V. (oka@dictum.ru), **Erekhinskaya T. N.** (te@dictum.ru), **Skatov D. S.** (ds@dictum.ru),
DICTUM Ltd., Nizhny Novgorod, Russia

The paper describes functional ambiguity of punctuation marks in the Russian language. A formal model of isolations and series of coordination members is presented. Mathematical target setting for punctuation use in syntax parsing and the algorithm for this task are suggested.

TRANSLATION OF GERMAN PARTICLE *DOCH* USED IN STATEMENTS INTO RUSSIAN (IN STATEMENTS): *VED'*, *ŽE*, *VSE ŽE* OR *VSE-TAKI*?

Orlova S. V. (svetlachok-star@yandex.ru), Lomonosov Moscow State University, Russia

The paper is devoted to the comparative analysis of the semantics of the German particle DOCH in statements and its translation equivalents taken from German-Russian dictionaries — the Russian particles VED', ŽE, VSE ŽE and VSE-TAKI.

ETYMOLOGICAL DICTIONARY: LEXICOGRAPHIC STRUCTURE AND REPRESENTATION IN DIGITAL ENVIRONMENT

Ostapova I. V. (iros@zeos.net), Ukrainian Linguo-Information Fond, National Academy of Sciences of Ukraine

A technology for building an instrumental system for supporting the dictionary in digital environment was developed. The technology is based on a formal model of lexicographic system of etymological dictionaries. The main focus is given to mechanisms of language indexation.

POSSESSIVES AND MANNER OF ACTION NOUNS: CORPUS BASED EXPLORATION

Paducheva E. V. (elena708@gmail.com), Institute of Scientific and Technical Information (VINITI), Russian Academy of Sciences

Possessives (i.e. possessive pronouns and adjectives) resemble the genitive, but possessive Subject co-occurs with a genitive Object in the context of a verbal noun (мейерхольдовская постановка Ревизора), while genitive Subjects are not compatible with genitive Objects. Possessive-genitive diathesis serves as a diagnostics for NOUNS OF MANNER.

DERIVATIONAL PATTERNS AND SYNTACTIC POSITIONS OF DEVERBAL NOMINALS (ON CORPUS DATA)

Pazelskaya A. G.(anna_pz@abbyy.com), ABBYY Software

This paper is a part of general study of differences in behaviour in Russian deverbal nominals derived via various patterns. The investigation is done on the basis of corpus data, mostly obtained from the Russian National Corpus. We study preferences of nominals ascending to the three most productive derivational patterns with respect to the syntactic position of the resulting nominal in a sentence.

PROSODY OF THE GERMAN VOCATIVE NPs IN CONTRAST TO THE RUSSIAN ONES

Palko M. L. (m_palko@mail.ru) Institute of Linguistics (Russian Academy of Sciences)

The prosody of the German vocative NPs is discussed as contrasted to the prosody of the Russian vocatives. The analysis shows that the German vocatives do not allow for prestressed articulations that are highly characteristic of the Russian vocatives used in unofficial and close contacts between the hearer and the listener, cf. MOLODOJ chelovek! with a wordform molodoj to be accented. The non-vocative NPs also demonstrate more restrictions in prestressed patterns formation, which seems to be the typological parameter of German and of most West European languages.

NONVERBAL COMMUNICATIVE ACT OF CONSOLATION: MATERIALS FOR A DICTIONARY OF NONVERBAL COMMUNICATIVE ACTS

Pereverzeva S. I. (P_Sveta@hotmail.com), Russian State University for the Humanities, Russia

The paper discusses some issues regarding the dictionaries of Russian speech acts and Russian nonverbal acts. I provide a preliminary draft of a dictionary entry "consolation" as an example of lexicographical description of nonverbal acts.

THE ROLE OF DISCOURSE MARKERS IN LOCAL DISCOURSE STRUCTURE: A CORPUS STUDY

Podlesskaya V. I. (podlesskaya@ocrus.ru), Russian State University for the Humanities
Kibrik A. A. (aakibrik@gmail.com), Institute of Linguistics RAS

Based on a corpus of spoken narratives, the study shows how discourse markers can be differently integrated into local discourse structure: some can be used as a separate "minimal discourse unit", while others are always integrated into a bigger unit with a propositional meaning. The two discourse markers most frequent in the corpus, VOT and NU, are compared and VOT is shown to be less integrated into prosodic, linear and hierarchic structure than NU.

LOMONOSOV CONCORDANCE — CONCEPT AND IMPLEMENTATION

Polyakov A. E. (pollex@mail.ru), NTC «Informregistr»
Bergelson M. B. (mirabergelson@gmail.com), Lomonosov Moscow State University, Russia
Pilshckov I. A. (pilshch@yandex.ru), IMK of Lomonosov Moscow State University

This paper qualifies the concepts and terminology relevant to the development of comprehensive digital Concordance to the texts of Lomonosov, and discusses the practical decisions which are necessary for the implementation of this lexicographical product. The concordance is based on the corpus of author's texts supplied with structural, philological and grammatical markup. We describe the technology we use to build the corpus and the concordance, the principles of corpus markup, and the structure of concordance vocabulary entries, as well as its application to linguistic research.

SYNTACTICALLY THE INVARIANT METHOD OF IDENTIFICATION OF SEMANTICS OF THE INFORMATION

Potapov M. V. (potapov_mv@rgrta.ryazan.ru), Ryazan State Radio Engineering University

In the report the description of practically approved method of an estimation of the semantic maintenance of the information streams based on statistic — a linguistic way of primary processing of the bit information and approaches of the theory of recognition of images contains at the analysis of multivariate attributes

UNSUPERVISED PARSING

Potemkin, S. (potemkin@philol.msu.ru), Philological Faculty, Moscow State University, Russia

A statistical approach to parsing of raw text is described. The parsing algorithm builds a projective dependency tree in quadratic time after training on an unannotated corpus.

MULTI-TIER MARKUP OF SPEECH CORPUS FOR HYBRID RUSSIAN TTS SYSTEM "VITALVOICE"

Prodan A. I. (prodan@speechpro.com), **Korolkov E. A.** (korolkov@speechpro.com), **Oparin I. V.** (ilya@speechpro.com),
Talanov A. O. (andre@speechpro.com), Speech Technology Center, Russia

The paper deals with the features of a system for multi-level markup of speech corpora. These corpora are used for the hybrid Russian TTS system "VitalVoice" developed at Speech Technology Center (STC). VitalVoice is basically a Unit Selection TTS system complemented with triphone inventory. The basic advantage of this approach is that it allows getting speech units from the speech corpus in a quick and efficient way. The database consists of interrelated levels of markup (phrases, intonation models, words, syllables, etc.). The levels of markup, their use in the TTS system and automatic markup checking are described in detail.

SEMANTIC-DERIVATIONAL MODELS OF POLYSEMOUS ADJECTIVES: METAPHOR, METONYMY AND THEIR INTERACTION

Rakhilina E. V. (rakhilina@gmail.com), Institute Of Russian Language, RAS
Karpova O. S. (o_k_inbox.ru), Russian State University for Humanities
Reznikova T. I. (tanja.reznikova@gmail.com), All-Russian Institute of Scientific and Technical Information, RAS

The paper reports on a project intended to provide a corpus-based description of semantic-derivational models for Russian adjectives. The research deals with high-frequency adjectives in the attributive use denoting the quality of a person or thing. We discuss basic metonymical and metaphorical patterns and analyze several non-regular shifts.

THE SO-CALLED: SEMANTIC ANALYSIS OF PARENTHETICAL METALINGUISTIC PHRASES

Rozina R. I. (rarozina@yandex.ru), Russian Language Institute n.a. V. V. Vinogradov, Russian Academy of Sciences

The paper is concerned with meaning and textual functions of a group of parenthetical phrases expressing the speaker's attitude to the manner of speech. It is argued that their function is to ensure the transition in the text between different styles, the relation between which changes in the course of time, and that the meaning of these phrases is extended in the way that might be regular.

AUTHORSHIP IDENTIFICATION WITH SUPPORT VECTOR MACHINE IN CASE OF TWO POSSIBLE ALTERNATIVES

Romanov Aleksandr S. (ras@ms.tusur.ru), **Mescheriakov Roman V.** (mrv@keva.tusur.ru),
Tomsk state university of control system and radioelectronics

Authorship identification problem is viewed as a classification task. The importance of resolving the binary authorship classification problem for authorship identification is justified. Description and results of authorship identification experiment with support vector machine in the case of two possible alternatives are given.

STRATEGIES OF DELIMITATION OF SYNTACTIC UNITS IN SPONTANEOUS SPEECH

Ryko A. I. (aryko@mail.ru), **Stepanova S. B.** (stsvet_2002@mail.ru), Saint Petersburg State University, Russia

The paper discusses methods of dividing spontaneous speech into syntactic units using the Corpus of Spoken Russian. We analyze individual strategies of experts who took part in the experiment, and examine connections between the boundaries of sentences and their final intonation.

ON ENCYCLOPAEDIC DATA IN AN APPLIED SEMANTIC DICTIONARY

Semenova S. Yu. (Sonya_sem@mail.ru), INION RAS, Russia

Inclusion of information on ontological realities into a semantic dictionary, which is a trend in modern lexicography, corresponds to ideas of cognitive science with its focus on the wholeness of the information perception process. The paper is concerned with the encyclopaedic data within the NLP-aimed semantic dictionary that has the rigid formats for lexical data representation. Encyclopaedic functions in the RUSLAN machine semantic dictionary are considered. Some ways of loading and enhancement of the functions are discussed. A number of words and lexical classes relevant to certain types of encyclopaedic data are considered.

AN ONTOLOGY-BASED APPROACH TO FACT EXTRACTION

Sidorova E. A. (lena@iis.nsk.su), **Kononenko I. S.**(irina_k@cn.ru), A. P. Ershov Institute of Informatics Systems, Russia

An approach is proposed to develop fact extraction technology applicable in information systems of various kinds. The approach makes use of the knowledge base including domain ontology, domain vocabulary, model for text segmentation, and fact extraction schemes that relate vocabulary items and lexical-syntactic constructions to ontology entities.

MODELS AND METHODS FOR THE ANALYSIS OF HIERARCHICALLY STRUCTURED TEXTS

Skatov D. S. (ds@dictum.ru), **Erekhinskaya T. N.** (te@dictum.ru), **Okatiev V. V.** (oka@dictum.ru),
DICTUM Ltd., Nizhny Novgorod, Russia

The analysis of hierarchically structured texts (laws, standards etc.) is discussed. An overview of developments in the domain are given. The developed models and methods for the analysis of hierarchically structured texts are described.

EXPERIENCE OF SYSTEMATIZING KNOWLEDGE AND INTERNET RESOURCES FOR A KNOWLEDGE PORTAL ON COMPUTATIONAL LINGUISTICS

Sokolova E. G. (minegot@rambler.ru), Russian State University for Humanities, Moscow
Kononenko I. S. (irina_k@cn.ru), **Zagorulko Yu. A.** (zagor@iis.nsk.su), A. P. Ershov Institute of Informatics Systems SB RAS, Novosibirsk

The paper describes an experience of systematizing knowledge and internet resources for a knowledge portal on computational linguistics. A composition and structure of objects of the portal, place of the portal among other catalogues on computational linguistics, an experience of development of bilingual vocabulary of terms on computational linguistics with using procedures of automatic extraction of terms from text are considered.

THE USE OF LEXICO-GRAMMATICAL DATABASES IN THE RUSSIAN DIALECTAL LEXICOGRAPHY

Aleksandra V. Ter-Avanesova (teravan@mail.ru), Institute of Russian Language of Russian Academy of Sciences, Moscow
Sergej A. Krylov (krylov-58@mail.ru), Institute of Oriental Studies of Russian Academy of Sciences, Moscow;
Institute of System Analysis of Russian Academy of Sciences, Moscow

The lexico-grammatical database (LGDB) for Russian folk dialects with two [o]-like phonemes that was built with the help of StarLing informational system is significantly enriched. It includes now the data on a Middle Russian dialect of the village Pustosha (Shatura district, Moscow region, and a LGDB for Vologda suburban dialects, including about 30 thousand word-forms that represent about 4500 lexemes. The kernel dialectal corpus (KDC) contains texts with partial lexico-grammatical tagging.

LEXICAL FUNCTIONS AND SEARCH ENGINE OPTIMIZATION (BASED ON WORDS WITH NUMERIC VALUES)

Timoshenko Svetlana (timoshenko@iitp.ru), **Leonid Cinman** (cinman@iitp.ru), Institute for Information Transmission Problems, Russian Academy of Sciences

To provide more precise web search we have developed a special option in the ETAP-3 multifunctional NLP environment. The search query consisting of two or three words has been supplemented with the values of certain lexical functions to generate an incomplete sentence which lacks only the numeral information. We expect that it may help in searching numeral data like "The height of the Pisa tower". The results of the experiment show that the search precision index in this domain of knowledge increases by 24 % on the average.

APPLYING LINGUISTIC SEMANTICS AND MACHINE LEARNING METHODS TO SEARCH PRECISION IMPROVEMENT IN SEARCH ENGINE "EXACTUS"

Tikhomirov I. A. (matandra@isa.ru), **Smirnov I. V.** (ivs@isa.ru) Institute for Systems Analysis of RAS, Moscow

The paper considers problems of using linguistic semantics and machine learning methods in the Exactus search engine. An experimental evaluation of search quality showed that these methods improve search precision and recall. Prospects of applying linguistic semantics and machine learning methods in search engines are discussed.

ON THE PROBLEM OF VARIABILITY OF IMPERATIVE ASPECTUAL FORMS

Trub V. M. (trub44@ukr.net)

The paper deals with the correlation between different aspectual forms of imperative verbs. We believe that one of the aims of semantic interpretation of inducements conveyed by different aspectual forms consists in the explication of semantic differences between them and the explanation of causes of irregularities reflected in the use of a form opposed to the default one.

KAK BY (lit. 'as if, like') AND KONKRETNO (lit. 'specifically')

Uryson E. V. (x-uryson@mtu-net.ru) Institute of Russian Language, Moscow

The semantics of Russian colloquial "parasitic" particles KAK BY (lit. 'as if, like') and KONKRETNO (lit. 'specifically') is described. The goal is to show that their emergence in the language is due to the lexical system of the language. KAK BY in its first meaning denotes similarity, and the words denoting similarity usually have a meaning denoting a set (a class). This is the way of "desemantization" of the conjunction KAK BY. The particle KONKRETNO develops its parasitic meaning by analogy with the word VOOBSHCHE ('in general'); the cause is that some meanings of KONKRETNO are antonyms to some meanings of VOOBSHCHE.

MEANINGS OF THE PREPOSITIONS "PO" AND "K" IN RUSSIAN: ENCODING OF ADJUNCTS AND SEMANTIC ROLES

Usacheva M. N. (mashastroeva@gmail.com), Lomonosov Moscow State University, Russia

This work is devoted to the application of the spatial meaning description method (developed primarily for Dagestani languages but claimed to be typologically universal: see [Ganenkov 2002, 2005[, [Mazurova 2007]) to Russian prepositions "po" and "k".

PROCESSING INITIAL-STRESS AND NON-INITIAL-STRESS WORDS IN SPOKEN-WORD RECOGNITION IN RUSSIAN

Fedorova O. V. (olga.fedorova@msu.ru), **Shavrygina A. S.** (shavrygina@gmail.com), Lomonosov Moscow State University, Russia

The data of an experimental investigation of spoken-word recognition in Russian are presented. Two experiments showed that word recognition and word recall are faster and better in initial-stress word than in non-initial-stress words. The results support the metrical segmentation theory.

EASTERN ARMENIAN NATIONAL CORPUS www.eanc.net

Khurshudian V. G. (vk@corpustechnologies.com), **Daniel M. A.** (misha.daniel@gmail.com),
Levonian D. V. (dl@renovacapital.com), **Plungian V. A.** (plungian@gmail.com), **Polyakov A. E.** (pollex@mail.ru),
Rubakov S. V. (rubakov@gmail.com), Corpus Technologies

Eastern Armenian National Corpus (EANC) is a comprehensive linguistic database of annotated texts in Eastern Armenian from the mid 19th century to the present. The EANC contains about 110 million tokens and is enhanced with a powerful search engine. EANC is available at www.eanc.net.

ZERO CATEGORIES IN UNIVERSAL GRAMMAR

Zimmerling A. V. (meinmat@yahoo.com) Moscow State University for the Humanities, MGGU

The paper discusses the status of zero categories in general syntax. The taxon 'pro' is not sufficient for tagging all covert pronouns in finite clauses. Moreover, the notion of 'discourse pro-drop languages' is not a valid tool in syntactic typology. Discourse-linked dropping of anaphoric pronouns, coreferent deletion and constraint on overt realization of pro-forms are different syntactic operations. More specifically, I am challenging some points in Holmberg's analysis of Finnish pro and claiming that 1-2 person pro-forms regularly display features different from 3rd person pronominal zeros. Finally, I am discussing the status of 'Mel'čuk's zeros', e.g. theta-role sensitive zero lexemes and proving for that theta-role sensitive zero pronouns with an Agentive value and theta-role neutral pro-forms may coexist in one and the same language.

STATISTICAL ANALYSIS AND CONTEXTUAL RULES OF HOMOGRAPH DISAMBIGUATION ON TEXT-TO-SPEECH SYNTHESIS

Tsirulnik Liliya I. (L.tsirulnik@newman.bas-net.by), United Institute of Informatics Problems, National Academy of Sciences of Belarus
Svetlana G. Barbuk (sviatos@tut.by), Minsk State Linguistic University, Belarus
Boris M. Lobanov (Lobanov@newman.bas-net.by), United Institute of Informatics Problems, National Academy of Sciences of Belarus

The rules of accent position location in the homographs based on the results of contextual and statistical analysis of scientific and artistic text corpora are described. The implementation of the developed rules in Russian TTS synthesis system "Multi-Phone" increase the degree of adequacy of sense understanding of synthesized speech.

AN ALGORITHM OF LINK SPAM DETECTION

Sharapov R. V. (info@vanta.ru), **Sharapova E. V.** (goldenstuff@mail.ru), Murom Institute of Vladimir State University

Approaches to detecting spam links on the basis of the analysis of page content are considered. We focus on the detection of advertisement (paid) links. Features of paid links are analyzed. The algorithm of detecting a spam link is given.

COMMUNICATIVES AND METHODS OF ITS DESCRIPTION

Sharonov I. A. (Igor_sharonov.@mail.ru), Russian State University for the Humanities

Short dialogical utterances with fixed and vague grammatical structure are analyzed. We call these utterances "communicatives" and focus on the main principles underlying the classification of such language forms and ways of their pragmatic and conversational analysis. To describe the functioning of a communicative in conversation we need to clarify their semantic, formal and discursive characteristics, which include: — communicative intention or emotional state; — what kind of speech act — direct or indirect — a communicative represent; — the source form of the communicative and the mode of its transposition into communicative; — the discursive boundaries with adjacent utterances; — standard intonation patterns and other phonetic characteristics of the communicative in speech.

VARIATION, CONTINUATION, AND SERIALITY OF JOKES: PROBLEMS OF DATABASE CONSTRUCTION

Shmeleva E. Y. (eshkind@mail.ru), **Shmelev A. D.** (smelev.alexei@gmail.com), Institute of Russian Language, Moscow

The paper deals with different kinds of joke variation and intertextual relations between jokes. We discuss such phenomena as realization of a joke, versions of a joke, continuation of a joke, modification of the original joke, addition to the original joke, series of jokes, joke cycle.

SYNTACTIC AMBIGUITY RESOLUTION: PRIMING AND SELF-PRIMING EFFECTS

Yudina M. (dietiefe@yandex.ru) (ABBYY, Moscow State University)
Fedorova O. (olga.fedorova@msu.ru) (Moscow State University)

The report is devoted to the first experimental research on the influence of syntactic priming on syntactic ambiguity resolution of relative clauses in Russian. Within the frame of syntactic priming we can see two effects: the syntactic priming itself and self-priming (persistent preference of subject's own syntactic strategy).

BEST RECOGNIZABLE WORDS UNDER DIFFERENT EXPERIMENTAL SETTINGS

Iagounova E. V. (iagounova_elena@mail.ru), St.Petersburg State University

Basic features of the sets formed by the words, best recognizable under white-noise masking and within meaningless text fragments have been analyzed. It is observed that the sets are crucially dependent on such broad text parameters as professional text vs. fiction and dynamic vs. static text.

STRUCTURING OF ATTRIBUTIVE WORD MEANINGS IN RUSSNET THESAURUS (IN RUSSIAN ADJECTIVES OF PERCEPTION)

Yavorskaya M. V. (yav.mas@gmail.com), **Azarova I. V.** (ivazarova@gmail.com), Saint Petersburg State University, Russia

Adjectives with perceptional meanings are described. We focus on the problem of attributive meanings structuring for computer thesaurus RussNet. 178 attributive word-meaning pairs are marked up in the random samples of corpus contexts. Attributes for different spheres of perception are compared.

RUSSIAN VOCATIVES: LEXICON AND CONSTRUCTIONS

Yanko T. E. (tanya_yanko@list.ru), Institute of Linguistics (Russian Academy of sciences)

According to Zwicky, semantically parallel NPs often have distinct vocative properties. Whether a given NP can be used as a call or an address is a dictionary information. In this paper a variety of specific vocative strategies and vocative constructions that change a vocative potential of lexical items is analyzed.

DEVELOPMENT AND IMPLEMENTATION OF MULTILINGUAL OBJECT TYPE TOPONYM-REFERENCED TEXT CORPORA FOR OPTIMIZING AUTOMATIC IMAGE DESCRIPTION GENERATION

Gornostay T. (tatjana.gornostaja@tilde.lv), Tilde, Riga (www.tilde.com),
Aker A. (a.aker@dcs.shef.ac.uk), Department of Computer Science, University of Sheffield

The fast growing amount of images available on the web has motivated development of automatic approaches for image description generation. Using multi-document summarization for this task has been proposed recently. This paper describes a method for developing and implementing object type toponym-referenced text corpora in the context of optimizing the multi-document summarization for generating toponym-referenced descriptions of images. Object type corpora are developed for four different languages: English, German, Italian and Latvian.

TRANSCRBING, STRUCTURING AND TEMPORAL ANALYSIS OF FLUENT SPEECH CORPUS FOR A UNIT SELECTION TTS SYSTEM FOR ESTONIAN

Mihkla M. (meelis@eki.ee), **Kiissel I.** (indrek@eki.ee), **Nurk T.** (tonis@eki.ee), **Piits L.** (liisi@eki.ee), Institute of the Estonian Language

The paper reports the development of a speech corpus for Estonian text-to-speech synthesis based on unit selection. The process of transforming an orthographic Estonian text into a pronounced text, requiring the consideration of quantity, palatalization and other essential features of an Estonian pronounced text, is described. In order to optimize the unit selection algorithm and to guarantee the necessary quality of the synthetic speech the whole speech database is represented as a phonological tree. We present the evidence that the collocational strength shortens the duration of words and that contextual predictability is a significant feature to be considered in developing models of word duration.

THE DYNAMICS OF ADJECTIVE MEANING

Partee Barbara H. (partee@linguist.umass.edu), University of Massachusetts, Amherst, MA, USA; Moscow State University

Meaning and context interact dynamically; how can one account for context-dependence without abandoning compositionality? We illustrate with the semantics of different kinds of adjectives. We show how compositional semantics sheds light on word meaning, and how compositional semantics, lexical semantics, and context all interact.

ONTOLOGICAL SEMANTICS AND ABDUCTION: PARSING ELLIPSIS

Petrenko M. (mpetrenk@gmail.com), Princess Dashkova Moscow Humanities Institute, Russia

New avenues for modeling abductive reasoning within the framework of Ontological Semantics are explored. Specifically, the rich knowledge resources and dynamic parsing module of Ontological Semantics allow processing elliptic input with a set of inference rules, which establish on the one hand, dependencies between verbalized and non-verbalized case-roles across clauses, and on the other hand, dependencies between scalar attribute values and specific event classes. Examples are provided to illustrate each case.

Авторский указатель

Научное издание

Компьютерная лингвистика и интеллектуальные технологии

По материалам ежегодной Международной конференции «Диалог»

Периодическое издание. Выпуск 9 (16). 2010

Ответственный за выпуск **Левченкова И. А.**
Вёрстка **Климентовский К. А.**

Издательский центр РГГУ
125993, Москва, Миусская пл., д. 6
Тел.: +7 (499) 973 42 00

Подписано в печать 12.05.2010 г.
Формат 60×84/8
Бумага офсетная.
Усл. печ. л. XX,X
Тираж 200 экз. Заказ №

Отпечатано с готового оригинал-макета в типографии
ООО «Издательско-полиграфический центр Маска»
117246, Москва, Научный пр-д, д. 20, стр. 9