Компьютерная лингвистика и интеллектуальные технологии

По материалам ежегодной Международной конференции «Диалог» (2015)

Выпуск 14

В двух томах

Том 1. Основная программа конференции

Computational Linguistics and Intellectual Technologies

Papers from the Annual International Conference “Dialogue” (2015)

Issue 14

Volume 1 of 2. Main conference program

УДК 80/81; 004
ББК 81.1
К63

Программный комитет конференции выражает
искреннюю благодарность Российскому фонду
фундаментальных исследований за финансовую поддержку,
грант № 15-07-20554 Г

Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Москва, 27–30 мая 2015 г.). Вып. 14 (21): В 2 т. Т. 1: Основная программа конференции. — М.: Изд-во РГГУ, 2015.

Сборник включает 69 докладов международной конференции по компьютерной лингвистике и интеллектуальным технологиям «Диалог 2015», 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.

Предисловие

14-й выпуск ежегодника «Компьютерная лингвистика и интеллектуальные технологии» содержит избранные материалы 21-й Международной конференции «Диалог». На основании мнений наших рецензентов для публикации в ежегоднике Редсоветом было отобрано 69 докладов из числа примерно ста работ, которые были рекомендованы по результатам рецензирования для представления на конференции в этом году.

Работы в сборнике отражают все основные направления исследований в области компьютерного моделирования и анализа естественного языка, представленные на конференции:

- Формальные модели языка и их применение в компьютерной лингвистике
- Модели и методы семантического анализа текста
- Лингвистические онтологии и извлечение знаний
- Теоретическая и компьютерная лексикография
- Методики тестирования технологий и верификации результатов лингвистических исследований (Dialogue Evaluation)
- Компьютерные лингвистические ресурсы и их связывание (Linked Data)
- Корпусная лингвистика: создание, разметка, методики применения и оценка корпусов
- Анализ Social Media
- Машинный перевод текста и речи
- Лингвистический анализ речи
- Модели общения. Коммуникация, диалог и речевой акт
- Мультимодальная лингвистика
- Компьютерный анализ документов: классификация, поиск, тематический анализ, оценка тональности и т. д.

В соответствии с традициями «Диалога», старейшей и крупнейшей конференции по компьютерной лингвистике в России, отбор работ основывается на представлении о важности соединения новых инженерных методов и технологий анализа языковых данных с результатами серьезных лингвистических исследований. Одной из важнейших целей конференции была и остается поддержка создания современных компьютерных ресурсов и технологий для русского языка.

В этом году продолжается и традиция проведения в рамках направления Dialogue Evaluation тестирований технологий решения отдельных задач компьютерного анализа русского языка. Значимость таких мероприятий трудно переоценить, поскольку их результаты создают основу для сравнительной оценки эффективности результатов в соответствующих областях исследований.

В этом году было проведено два таких тестирования: сравнивались различные подходы к анализу т.н. аспектного сентимента и оценке семантической близости слов.

Sentiment Analysis является важным самостоятельным прикладным направлением компьютерной лингвистики, особенно в той постановке, которая была предложена участникам: не только определение общей тональности документа, но и выделение и оценка в нем отдельных аспектов выражения мнения.

Тестирование методов определения семантической близости слов является важным для понимания сложной картины в современной вычислительной семантике, где конкурируют и взаимодействуют традиционные словарные и новые дистрибуционные методы исследования лексических значений.

Наиболее значимые работы, представленные участниками этих тестирований, выделены в отдельный второй том ежегодника. Там же опубликованы и итоговые статьи организаторов.

Программный комитет конференции выражает особую признательность Наталье Лукашевич и Александру Панченко за особую роль в организации и проведении этих тестирований.

Среди особых направлений «Диалога» в этом году — исследования в области русского мультимодального дискурса. Интерес к языковой коммуникации как целому всегда был характерным для нашей конференции, выросшей, напомним, из семинара, носившего название «Модели общения». Доклады мультимодального направления составляют важную часть этого сборника.

Статьи в сборнике публикуются на русском и английском языках. При выборе языка публикации действует следующее правило:

- доклады по компьютерной лингвистике должны подаваться на английском языке. Это расширяет их аудиторию и позволяет привлекать к рецензированию международных экспертов.

- доклады, посвященные лингвистическому анализу русского языка, предполагающие знание этого языка у читателя, подаются на русском языке (с обязательной аннотацией на английском языке).

Несмотря на традиционную широту тематики представленных на конференции и отобранных в сборник докладов они не могут дать полной картины направлений «Диалога». Ее можно получить с помощью сайта конференции www.dialog-21.ru, на котором представлены обширные электронные архивы «Диалогов» последних лет и все результаты проведенных тестирований.

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.

Учредителями конференции являются:

- Институт лингвистики РГГУ
- Институт проблем информатики РАН
- Институт проблем передачи информации РАН
- Компания ABBYY
- Компания Yandex
- Филологический факультет МГУ

Конференция проводится при поддержке Российской ассоциации искусственного интеллекта.

Международный программный комитет

Байтин Алексей Владимирович	Компания Yandex
Богуславский Игорь Михайлович	Политехнический университет Мадрида
Буате Кристиан	Гренобльский университет
Гельбух Александр Феликсович	Национальный политехнический институт, Мехико
Иомдин Леонид Лейбович	Институт проблем передачи информации РАН
Кобозева Ирина Михайловна	Филологический факультет МГУ
Козеренко Елена Борисовна	Институт проблем информатики РАН
Корбетт Гревил	University of Surrey, UK
Кронгауз Максим Анисимович	Институт Лингвистики РГГУ
Лукашевич Наталья Валентиновна	НИВЦ МГУ
Маккарти Диана	Lexical Computing Ltd., UK
Мельчук Игорь Александрович	Монреальский университет
Ниврe Йоаким	Уппсальский университет
Ниренбург Сергей	Университет Нью-Мексико
Осипов Геннадий Семёнович	Институт программных систем РАН
Попов Эдуард Викторович	РосНИИ информационной техники и САПР
Раскин Виктор	Purdue University, USA
Селегей Владимир Павлович	Компания ABBYY
Хови Эдуард	University of Southern California
Шаров Сергей Александрович	University of Leeds, UK

Организационный комитет

Селегей Владимир Павлович, *председатель*	Компания ABBYY
Байтин Алексей Владимирович	Компания Yandex
Беликов Владимир Иванович	Институт русского языка им. В. В. Виноградова РАН
Браславский Павел Исаакович	Kontur Labs; Уральский федеральный университет
Добров Борис Викторович	НИВЦ МГУ
Иомдин Леонид Лейбович	Институт проблем передачи информации РАН
Кобозева Ирина Михайловна	Филологический факультет МГУ
Козеренко Елена Борисовна	Институт проблем информатики РАН
Лауфер Наталия Исаевна	ООО «проФан Продакшн»
Ляшевская Ольга Николаевна	Universitet i Tromsø, Norway
Сердюков Павел Викторович	Компания Yandex
Соколова Елена Григорьевна	РосНИИ искусственного интеллекта
Толдова Светлана Юрьевна	Филологический факультет МГУ
Шаров Сергей Александрович	University of Leeds, UK

Секретариат

Белкина Александра Андреевна, *секретарь оргкомитета*	Компания ABBYY
Атясова Анастасия Леонидовна, *координатор*	Компания ABBYY

Рецензенты

Августинова Таня
Азарова Ирина Владимировна
Апресян Валентина Юрьевна
Байтин Алексей Владимирович
Баранов Анатолий Николаевич
Беликов Владимир Иванович
Богданов Алексей Владимирович
Богданова-Бегларян Наталья Викторовна
Богуславский Игорь Михайлович
Бонч-Осмоловская Анастасия Александровна
Браславский Павел Исаакович
Васильев Виталий Геннадьевич
Гельбух Александр Феликсович
Гершман Анатолий
Гращенков Павел Валерьевич
Гриненко Михаил Михайлович
Гришина Елена Александровна
Губин Максим Вадимович
Даниэль Михаил Александрович
Добров Борис Викторович
Добровольский Дмитрий Олегович
Добрынин Владимир Юрьевич
Зализняк Анна Андреевна
Захаров Виктор Павлович
Захаров Леонид Михайлович
Иванов Владимир Владимирович
Иомдин Борис Леонидович
Иомдин Леонид Лейбович
Кибрик Андрей Александрович
Кобозева Ирина Михайловна
Козеренко Елена Борисовна
Коротаев Николай Алексеевич
Котельников Евгений Вячеславович
Котов Артемий Александрович
Крейдлин Григорий Ефимович
Кронгауз Максим Анисимович
Леонтьев Алексей Петрович
Лобанов Борис Мефодьевич
Лукашевич Наталья Валентиновна
Ляшевская Ольга Николаевна
Маккарти Диана
Минлос Филипп Робертович
Недолужко Анна Юрьевна
Новицкий Валерий Игоревич
Пазельская Анна Германовна
Панченко Александр Иванович
Паперно Денис Аронович
Пиперски Александр Чедович
Подлесская Вера Исааковна
Савельев Василий Евгеньевич
Селегей Владимир Павлович
Смирнов Иван Валентинович
Сокирко Алексей Викторович
Соколова Елена Григорьевна
Сорокин Алексей Андреевич
Старостин Анатолий Сергеевич
Тихомиров Илья Александрович
Толдова Светлана Юрьевна
Турдаков Денис Юрьевич
Урысон Елена Владимировна
Федорова Ольга Викторовна
Хови Эдуард
Хорошевский Владимир Федорович
Циммерлинг Антон Владимирович
Шаров Сергей Александрович
Янко Татьяна Евгеньевна

Contents*

Основная программа конференции

* The reports of each section are ordered by the surname of the first author in compliance with the English alphabet.

Раздел I.

Основная программа конференции

СВЯЗЬ СЕМАНТИЧЕСКИХ И КОММУНИКАТИВНЫХ СВОЙСТВ ЯЗЫКОВЫХ ЕДИНИЦ

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

Национальный исследовательский университет Высшая школа экономики; Институт русского языка им. В. В. Виноградова РАН, Москва, Россия

Цель данной работы — определить, наличие каких семантических компонентов в значении способствует тому, чтобы языковая единица приобретала лексикализованные просодические и коммуникативные свойства, в частности, обязательное акцентное выделение и способность формировать самостоятельную рему или контрастную тему. В работе показывается, что акцентные и коммуникативные свойства лексем коррелируют с их семантическими свойствами. В частности, на примере сравнения различных значений слов *только, правда, еще, вообще, по крайней мере* и некоторых других, а также сравнения семантически близких языковых единиц *хотя, несмотря на, пусть*, демонстрируется связь способности нести на себе акцентное выделение и быть фокусом внимания в высказывании в у единиц, содержащих семантические компоненты противопоставления, добавления и высокой степени. С другой стороны, наличие в значении уступительного компонента, как правило, ограничивает способность лексемы к акцентному выделению и коммуникативной самостоятельности. Сформулированные тенденции подтверждаются данными мультимедийного корпуса НКРЯ.

Ключевые слова: семантика, прагматика, тема, рема, пресуппозиция, ассерция, акцентное выделение, противопоставление, высокая степень

CORRELATION BETWEEN SEMANTIC AND COMMUNICATIVE PROPERTIES OF WORDS

Apresjan V. Ju. (valentina.apresjan@gmail.com)

National Research University Higher School of Economics; Vinogradov Russian Language Institute of the Russian Academy of Sciences, Moscow, Russia

The objective of this paper is to determine what semantic components in the meaning of a word facilitate its lexicalization as prosodically marked and aid its focalization in an utterance. The paper demonstrates that prosodic and communicative properties of a word correlate with its semantic properties.

In particular, a case study of different senses of the words *tol'ko* 'only', *pravda* 'true', *eshche* 'still, more', *voobshche* 'in principle, generally', *po krajnej mere* 'at least' and some others reveals that focalization and prosodic marking in a word are triggered by the semantics of contrast, high degree, and addition. On the other hand, semantics of concession in the meaning of a word limits its ability for accentual marking and focalization. The observed correlations between semantics/on the one hand, and prosody and communicative properties, on the other, are confirmed by the multimedia corpus data.

Key words: semantics, pragmatics, topic, focus, presupposition, assertion, accentual marking, opposition, high degree

Введение[1]

Известно, что есть семантические компоненты, которые тяготеют к той или иной семантической или коммуникативной части высказывания: например, отрицание — типичный ассертивно-рематический компонент, в то время как указание на существование часто входит в пресуппозицию и тему. Соотношение ассертивно-пресуппозитивного и темо-рематического членения высказывания — отдельная проблема, которая не рассматривается в рамках данной работы [см. Partee 1993]. Задача данной работы — определить, наличие каких семантических компонентов в значении способствует тому, чтобы языковая единица приобретала способность формировать самостоятельную рему или же контрастную тему — т. е., такие части высказывания, которые привлекают к себе наибольшее внимание слушающего. Очевидно, что помимо семантических факторов, в этом играет роль еще и просодия — языковая единица должна быть полноценным фонетическим словом для того, чтобы иметь способность нести на себе интонационное ударение. Примером приобретения частицей способности нести акцентное выделение и, соответственно, формировать самостоятельную рему, при редупликации служат частицы *чуть, еле, едва*. Однако предметом данной работы являются семантические обстоятельства, способствующие появлению у языковой единицы определенных коммуникативных свойств — а именно, способности становиться фокусом внимания.

Интересный материал в этом смысле представляют собой слова, у которых за разными значениями закреплены разные коммуникативные свойства.

[1] Исследование осуществлено при поддержке следующих грантов: гранта РГНФ №13-04-00307а «Подготовка второго выпуска Активного словаря русского языка» (2013–2015), гранта НШ-3899.2014.6 для поддержки научных исследований, проводимых ведущими научными школами РФ «Разработка материалов для Активного словаря русского языка» (2014–2015), гранта Программы фундаментальных исследований Президиума РАН «Историческая память и российская идентичность» — «Основной лексический фонд русского языка как элемент русской культуры: системная организация лексики и ее отражение в словаре». Кроме того, в данной научной работе использованы результаты проекта «Корпусные технологии в лингвистических и междисциплинарных исследованиях», выполненного в рамках Программы фундаментальных исследований НИУ ВШЭ в 2015 году.

Лексикализация акцентных выделений — известное явление, описанное в работах [Николаева 1985, Апресян 1995, Богуславский 1985]. Изучение разных значений слов, обладающих разными лексикализованными коммуникативными свойствами, может пролить свет на то, как семантика языковой единицы связана с ее функцией в коммуникативной структуре высказывания.

Разные коммуникативные свойства отмечаются у разных значений частиц *еще, вот, вон* [Николаева 1985, Николаева 2000, Апресян 1995, Богуславский 1985], прилагательного *один* ([Николаева 1985, Николаева 2000, Богуславский 1985], прилагательного *настоящий* [Апресян 1995] и некоторых других. Корреляция семантических свойств лексем с их акцентными выделениями рассматривается в работе [Апресян 1995: 190–191] на примере слов *еще, должен* и *мочь*, *вообще*, а также вводных оборотов со значением высокой степени достоверности типа *разумеется, конечно*. В работе [Булыгина, Шмелев 1997] рассматриваются с этой точки зрения разные значения наречия *немного*. Представляется, что этот материал можно расширить. В данной работе рассматриваются разные значения вокабул *по крайней мере, кроме, правда, только*, обладающие разными просодическими и коммуникативными свойствами. Кроме того, рассматриваются коммуникативные различия между семантически близкими единицами, которые также помогают пролить свет на причины, способствующие развитию у них способности привлекать к себе акцентное выделение и, соответственно, формировать самостоятельную рему или контрастную тему в высказывании.

В работе используются данные Мультимедийного корпуса НКРЯ, позволяющие верифицировать интуиции исследователя относительно просодических особенностей изучаемых языковых единиц. Акцентным выделением мы называем такое произнесение языковой единицы, когда она в большей степени маркирована интонационно, чем все ее соседи в пределах синтаксической составляющей. Чаще всего речь идет о фразовом ударении, т. е. случаях, когда языковая единица просодически выделяется относительно всего высказывания и, соответственно, формирует самостоятельную рему или является интонационным центром ремы. Ср. пример из МУРКО (стрелочкой обозначено интонационно выделенное слово): *Э́то по-ви́димому да́же не заи́мствования/во́т наско́лько мы́ смогли́ поня́ть/по-англи́йски ничего́ тако́го сло́во «center» не зна́чит/ви́димо/тру́дно предположи́ть заи́мствование из како́го-то ↓ещё языка́ зде́сь* [Б. Иомдин. Доклад на конференции «Диалог 2013» (2013)], где *еще* маркировано относительно всей фразы *ви́димо/тру́дно предположи́ть заи́мствование из како́го-то ↓ещё языка́ зде́сь*. Случаи, где исследуемая языковая единица маркируется наравне с другой и, соответственно, не образует самостоятельной ремы или другой самостоятельной коммуникативной единицы (например, контрастной темы), а лишь является их равноправной частью, не считаются акцентно выделенными. Ср. следующую фразу из МУРКО: *Или во́т ↓ещё ↓приме́р* [Б. Иомдин. Доклад на конференции «Диалог 2013» (2013)], где слова *еще* и *пример* равноударны, т. е. *еще* не обладает коммуникативной самостоятельностью.

В качестве акцентного выделения также рассматривается интонационное маркирование контрастной темы, которое, хотя и не является единственным в высказывании, однако однозначно выделяет языковую единицу в пределах

ее синтаксической составляющей; ср. пример из МУРКО *Скажи́те/а каки́е-нибудь помеще́ния вот… ↓кро́ме прихо́жей ту́т ↓е́сть?* [Олег Куваев. Масяня, м/ф (2002–2008)], где одинаково ярко маркированы контрастная тема (*кроме*; слово *прихожей* безударно) и рема (*есть*).

1. Противопоставление

Семантическая идея противопоставления способствует развитию у языковой единицы рематических свойств, что неудивительно. Противопоставление возможно по отношению к чьим-либо ожиданиям, которые в таком случае формируют тему, а противоречащая ожиданиям информация естественным образом становится ремой. В качестве примера можно привести разные значения *по крайней мере*, *правда*, *только* и *вообще*.

1.1. *По крайней мере* в разных значениях

В работе [В. Апресян 2006] выделяются количественное и уступительное значения частицы *по крайней мере.* В количественном значении *по крайней мере* используется для выражения уверенности говорящего в том, что количество P составляет не менее, чем Q: *Молодой человек, я был в этом городе, по крайней мере, одиннадцать раз* (В. Аксенов); *Когда я зашел в свою комнатенку, то насчитал по крайней мере четырнадцать предметов (бытового назначения), с помощью которых можно отразить атаку взвода автоматчиков* (А. Азольский); *Если бы на рукаве были звёзды, а на отложном воротнике ромбы, то его можно было бы принять по крайней мере за молодого комбрига или даже начдива* (В. П. Катаев).

Количественная частица *по крайней мере 1* имеет следующее значение:

(1) *P по крайней мере 1 Q* (*На заседании присутствовало по крайней мере тридцать человек*) = ‘говорящий не знает точное значение P; говорящий утверждает, что P не меньше чем Q; говорящий утверждает, что P —это не мало’.

В этом значении частица *по крайней мере 1* является монотонно-возрастающей, т. е. из фразы вида *На заседании присутствовало по крайней мере двадцать человек* следует, что на заседании присутствовало любое количество человек, которое включается в двадцать — 19 человек, 18 человек и пр. Как видно из толкования, говорящий уверен в своем определении нижней границы P и, более того, считает P достаточно значительным (часто в противиположность чьим-то ожиданиям). Все эти компоненты значения — противоречие ожиданиям, уверенность, значительность [ср. Апресян 1995:191] способствуют развитию у этой частицы способности быть фокусом внимания в предложении. В этом значении *по крайней мере* может сопровождаься акцентным выделением: *Переписывать экзамен придется пяти ученикам по ↓крайней мере*, формируя интонационный

центр ремы. Данные МУРКО подтверждают это предположение: из 16 примеров на частицу *по крайней мере* в контексте числительного, относящихся к значению ‘не меньше’, в четырех *по крайней мере* имеет самостоятельное акцентное выделение. Интересно, что интонационно маркированы примеры, где речь идет о достаточно больших цифрах; контексты с малыми числительными такого эффекта не обнаруживают: *Вот э́то и е́сть та́ зага́дка/над кото́рой бьётся уже́/ по ↓кра́йней мере/сто́ пятьдеся́т ле́т мно́жество разли́чных вот учёных* [Асимметрия и возникновение жизни. Программа «Гордон» (НТВ) (2003)]; *Есте́ственно/ сра́зу спра́шивается/а что́ ну́жно? Ну́жно вот де́льта э́ль к э́ль/по ↓кра́йней ме́ре/ де́сять в ми́нус два́дцать оди́н/лу́чше в ми́нус два́дцать два́* [Гравитационные волны. Программа «Гордон» (НТВ) (2003)]; *Та́м в ка́ждом подъе́зде по ↓кра́йней ме́ре два́дцать кварти́р* [Павел Арсенов, Кир Булычев. Гостья из будущего, к/ф (1984)].

В своем уступительном значении *по крайней мере 2* предполагает монотонное убывание. Сравнение идет с чем-то большим, а не с чем-то меньшим, в отличие от *по крайней мере 1*: *Если пьющему не можешь помочь, то, по крайней мере, не мешай ему* (Ф. Искандер); *К великому сожалению подростка Соловьева, там не было ни реки, ни, по крайней мере, пруда* (Е. Водолазкин); *Хоть, по крайней мере, не мелькай перед глазами, ступай в угол и молчи в тряпочку* (В. П. Катаев). Из фразы вида *По крайней мере пруд-то там есть*?, следует отсутствие там любого большего, чем пруд, водоема. Прагматически *по крайней мере 2* указывает на неуверенность говорящего: говорящий не только уверен в том, что желательная для него ситуация P не будет иметь места, но даже не знает, будет ли иметь место меньшая ситуация Q:

(2) *Если уж не2 P, то по крайней мере Q* = ‘субъект или говорящий очень хочет, чтобы было P; субъект или говорящий считает, что P не будет; субъект или говорящий считает, что может быть Q; субъект или говорящий хочет, соглашается, или делает или готов делать так, чтобы было Q, которое мало или меньше P’.

Поэтому для *по крайней мере 2* нехарактерно акцентное выделение и функция самостоятельной ремы или интонационного центра ремы: ??*Если ты не сможешь прийти, то хоть позвони по ↓крайней мере*. Обычно интонационно маркируется именная или глагольная группа, к которой относится *по крайней мере*, т. е. по *крайней мере 2* не образует самостоятельной коммуникативной единицы. Ср. примеры из МУРКО, где *по крайней мере* употребляется в контексте *хоть*, т. е. речь явно идет о частице *по крайней мере 2*; во всех примерах *по крайней мере* не имеет акцентного выделения: *Ну ведь я́ бы хо́ть ↓помечта́л бы по кра́йней мере!* [Мать Бальзаминова, Людмила Шагалова, жен, 41, 1923]; *За Ратми́ром. … за Ратми́ром! О́н/по кра́йней ме́ре/хоть ↓ла́ять уме́ет* [Михаил Ильенко. Каждый

---

[2] В сентенциальном входе для введения валентности P используется конверсив лексем *по крайней мере 2*, хотя бы 1, хоть 3 — уступительная конструкция *если уж не…то*, — в комбинации с которой они часто употребляются.

охотник желает знать…, к/ф (1985)]; *Я́/по кра́йней мере/хо́ть ↓откро́венен* [Николай Губенко. Подранки, к/ф (1977)]; *А скажи́те/по кра́йней ме́ре/хо́ть ка́к о́н ↓вы́глядит/э́тот Родио́нов… ва́ш?* [Сергей Герасимов и др. Непобедимые, к/ф (1942)].

1.2. *Правда* в разных значениях

Похожие акцентно-коммуникативные свойства имеет наречие *правда* в значении ‘в самом деле’ (обозначим его *правда 1*): *Он и ↓правда об этом пожалел*; *Я ↓правда тебя люблю*; *И я сказал — покажи. У него ↓правда были пятьдесят баксов* (А. Геласимов); *Ах, у меня память отшибло! Хотя… Беда-то в том, что её у меня и ↓правда отшибло* (В. Белоусова). Для него предлагается следующее толкование:

(3) *Правда А1* ‘Говорящий утверждает, что ситуация А1 имеет место; до этого кто-то сомневался, что А1 имеет место’.

Правда 1, в отличие от *по крайней мере 1*, предполагает не потенциального, а реального оппонента, и поэтому степень противопоставления ожиданиям и настойчивость говорящего в утверждении своего мнения еще выше. Соответственно, *правда 1* не просто позволяет рематизацию, но **требует** ее: эта частица всегда произносится с акцентным выделением.

В своем уступительном значении *правда 2* не имеет акцентного выделения: *Катерина — родная сестра хозяина — приехала погостить из далёкой Сибири, правда ненадолго, проездом* (Б. Екимов).

Выборка из МУРКО подтверждает сформулированные просодические различия между разными значениями вокабулы *правда*. В пятидесяти проанализированных примерах значения распределились следующим образом: *правда* как существительное или в составе фразем (11 употреблений), *правда* как уступительный союз (21 употребление), *правда* в значении подтверждения вида *Я тебя правда люблю* (10 употреблений), *правда* в значении просьбы о подтверждении — *Я тебя люблю. — Правда?* (8 употреблений). Таким образом, численно преобладают интересующие нас служебные значения слова *правда*, в которых уступительное значение и значение подтверждения (с двумя подзначениями) имеют приблизительно одинаковое частотное распределение. Просодически они охарактеризованы следующим образом: все уступительные употребления безударны (интонационно выделяется опорное слово в пропозиции, к которой семантически относится *правда*), все употребления в значении подтверждения просодически выделены. Ср. *Инстру́кция по э́тому по́воду/во́т/благодаря́ фо́руму была́ напеча́тана/пра́вда/электро́нно* [Е. Г. Соколова, С. Ю. Толдова. Доклад на конференции «Диалог 2012» (2012)]; *Что́ же получа́ется? Органи́зм — одноклеточный/пра́вда — име́ет генети́ческую програ́мму со́бственного уби́йства* [Владимир Скулачев. Homo Sapiens Liberatus: человек, освобожденный от тирании генома. Проект Academia (ГТРК Культура) (2010)] (уступительное *правда*) vs. *О́й… Боле́знь-то и пра́вда зара́зная* [Олег Куваев. Масяня, м/ф (2002–2008)]; *Не́т.*

Не́т/спаси́бо большо́е/но мне́ пра́вда лу́чше в Пу́шкино! [Владимир Бортко, Александр Червинский. Блондинка за углом, к/ф (1984)] (*правда*-подтверждение).

Просодические различия связаны как с семантикой, так и с синтаксическими особенностями разных значений *правда*; так, в значении подтверждения *правда* часто формирует самостоятельную реплику, что требует ударности, а в уступительном значении имеет союзные свойства, что, напротив, способствует безударности. Однако союзность в принципе не исключает просодической выделенности; ср.: *Ставить свои условия будешь, когда тебя возьмут на работу.* ↓*Если тебя возьмут.* С другой стороны *правда* в значении подтверждения не всегда употребляется в качестве самостоятельной реплики. Так что различные просодические свойства уступительного *правда* и *правда*-подтверждения объясняются в первую очередь перечисленными различиями в семантике.

В ситуациях, когда контекст не дает достаточных оснований для различения между уступительным значением и значением подтверждения, интонация является единственным способом смыслоразличения (в данном примере — уступительное значение): Ра́зные оформле́ния таки́х сло́в/поня́тно/что норматѝвный вариа́нт — э́то во́т дефи́с/ту́т пра́вда ↓интере́сно/что поря́док мо́жет бы́ть ра́зный/«пла́тье-руба́шка»/«руба́шка-пла́тье» [Б. Иомдин. Доклад на конференции «Диалог 2013» (2013)].

Аналогично частице *правда 1* устроены частицы *действительно* и *в самом деле* со значением **уверенного подтверждения**: *Многие говорили, что они похожи: Снежка и Кямал. И* ↓*действительно что-то было* (В. Токарева); *— Знать, видел я её название где-то в расписании, мне думалось, выдумал, а она — вот она, и* ↓*в самом деле есть в наличности* (В. Астафьев).

У частицы *на самом деле* идея противопоставления выражена еще сильнее. Она имеет значение утверждения чего-либо, **противоположного** мнению слушающего [Баранов, Плунгян, Рахилина 1993]: *Он думает: «Как бы я хотел быть моряком». А на самом деле он не хочет быть никаким моряком* (Е. Е. Гришковец). Для нее предлагается следующее толкование:

(4) *На самом деле A1* ‘Говорящий утверждает, что ситуация A1 имеет место; до этого кто-то считал или утверждал, что A1 не имеет места’.

Соответственно, она не только рематична и требует акцентного выделения, но часто произносится с контрастным повышением тона. Как показывает выборка из МУРКО, *на самом деле* произносится с повышением тона на *самом* более, чем в трети случаев (20 примеров из 50 проанализированных). Прочие употребления также акцентно выделены; произнесения с выделением ↓*деле* и равноударные произнесения *на* ↓*самом* ↓*деле* делятся примерно в равной пропорции.

1.3. *Только* в разных значениях

В своем кванторном значении, описанном в работе [Богуславский 1985], *только 1* предполагает противопоставление одного элемента множества,

который обладает некоторым свойством, всем остальным, которые этим свойством не обладают: *В тот момент меня интересовала Сонька и только Сонька* (В. Белоусова). В противопоставлении в принципе принимают участие два элемента — единственный объект, обладающий свойством, и все остальные, которые им не обладают. *Только* вводит указание и фокусирует внимание на том единственном объекте, который обладает некоторым свойством; при этом все остальные объекты множества могут не упоминаться эксплицитно, поскольку в пресуппозицию *только* входит указание на то, что они этим свойством не обладают. Фразы типа *Только Вася решил задачу* содержат имплицитное указание на то, что никто другой ее не решил. Таким образом, ассерция (единственный объект обладает некоторым свойством) противопоставляется пресуппозиции (никто другой им не обладает), что способствует рематизации ассертивного элемента и его акцентному выделению. Кроме того, *только 1* содержит указание на количественную оценку говорящего, т. е. на модальную рамку: единственный объект, обладающим некоторым свойством, оценивается говорящим как малое количество, в противоположность возможным ожиданиям. *Только 1* в кванторном значении не требует акцентного выделения, но допускает его, поскольку обладает достаточным семантическим потенциалом для привлечения к себе внимания (противопоставление ассерции и модальной рамки ожиданиям, содержащимся в пресуппозиции): *Володька был эгоистичен в любви. Думал* ↓*только о себе, как солист. Один, и главный, и все должны под него подстраиваться* (В. Токарева).

В своем уступительном значении *только 2* не допускает акцентного выделения: *Сам он не ощущал голода, только хотелось пить* (В. Быков).

Для уступительного союза *только 2* и близкой ему по значению единицы *правда 2* предлагается следующее толкование:

(5) *P, правда <вот только>, Q* = ‘имеет место желательная ситуация P;
имеет место нежелательная ситуация Q; говорящий считает,
что в данной ситуации P несколько важнее, чем Q; говорящий
признает, что ситуация Q тоже может быть важной’.

Правда 2 и *только 2* прагматически «слабые» единицы: они вводят указание на менее важную ситуацию. Поэтому они не могут нести акцентное выделение и формировать рему.

Сформулированные корреляции между семантикой и просодией разных лексем *только* подтверждаются анализом употреблений по корпусу МУРКО. Из 26 вхождений *только*, встретившихся по запросу *только* bcomma, 15 представляют собой уступительное значение. Все они, за вычетом одного примера, интонационно безударны; маркировано только одно уступительное *только*, причем в театральной, замедленной речи: *— Э́э… ведь я́ живу́ недалеко́ от ва́с. — Неуже́ли? — В одно́м до́ме,* ↓*то́лько по друго́й ле́стнице* [А. Н. Островский. Бешеные деньги (постановка Малого театра) (1948)]. Из такого же количества проанализированных примеров употребления кванторного *только* акцентное выделение присутствует в девяти; ср. *Поня́тно почему́/«води́ть за́ нос» ока́зывается «имперфекти́ва та́нтум»/и «провёл за́ нос» появля́ется кра́йне ре́дко и* ↓*то́лько в конте́кстах/*

сня́той утверди́тельности [Д. О. Добровольский. Доклад на конференции «Диалог 2013» (2013)]; *Соотве́тственно э́э из приме́ров зде́сь ви́дно/что́ в ко́ми оно́ опи́сывается прилага́тельным/кото́рое покрыва́ет* ↓*то́лько фре́йм ко́лющих инструме́нтов* [М. Кюсева. Доклад на конференции «Диалог 2013» (2013)]; *И осо́бенность э́тих эгоце́нтриков в то́м/что́ (..) говоря́щий у ни́х появля́ется* ↓*то́лько в канони́ческой коммуникати́вной ситуа́ции/то́ есть в ситуа́ции диало́га* [Е. В. Падучева. Доклад на конференции «Диалог 2013» (2013)] и другие примеры.

1.4. *Кроме* в разных значениях

Интересный конраст к кванторному *только 1* представляет собой предлог *кроме 1*, который имеет очень похожее значение: *Он думает только о себе* ≅ *Он не думает ни о ком, кроме себя.* Однако если *только 1* позволяет акцентное выделение, то для *кроме 1* это невозможно: **Он не думает ни о ком,* ↓*кроме себя.* Это связано с иным фокусом внимания у предлога *кроме 1*, а также с отсутствием у этого предлога модальной рамки. Во-первых, *кроме 1* не предполагает малой количественной оценки; во-вторых, *кроме 1* предполагает два фокуса внимания — объект, являющийся исключением, и все остальные объекты. Для *кроме 1* обязательна реализация валентности всех остальных объектов, которая редко реализуется у *только 1*; ср. невозможность: **Кроме Васи не решил задачу.* Поэтому если у *только 1* фокус внимания на единственном объекте, у *кроме 1* он двойной, и единственный объект оказывается не в состоянии полностью «перетянуть» внимание и, соответственно, акцентное выделение на себя. Тем более, валентность прочих элементов множества реализуется *у кроме 1* либо отрицательным квантором (*никто, ничто*), либо квантором всеобщности (*все, весь*), т. е. сильно рематическими элементами, на которые, как правило, и падает акцентное выделение.

Однако в своем другом значении, значении добавления, где *кроме 2* синонимичен предлогу *помимо* (*Кроме французского, она знает английский и немецкий*), он обладает способностью к акцентному выделению. Ср. ↓*Кроме Васи, кто-то решил эту задачу?*

Данные корпуса МУРКО подтверждают эту гипотезу. Было 50 проанализировано примеров употребления *кроме*, из которых в 28 было представлено значение ‘исключения’, а в 22 — значение ‘добавления’. При этом в значении ‘исключения’ акцентное выделение присутствовало в одном примере, а в значении ‘добавления’ — в восьми: *Де́ло в то́м/что ми́р писа́лся двоя́ко до револю́ции — через «и́» восьмери́чное/кото́рое вы́ ви́дите сейча́с на экра́не/и «і» десятери́чное. «і» десятери́чное обознача́ло все́ значе́ния/*↓*кро́ме анто́нима сло́ва «война́»* [Рождение художественного текста. Программа «Гордон» (НТВ) (2003)] (*кроме* ‘исключения’) vs. ↓*Кро́ме бо́га Творца́/пу́сть и во мно́гих ли́цах/но в одно́й су́щности/египтя́не сло́вом «н э́чер» обознача́ли многочи́сленных ду́хов* [Боги Древнего Египта. Программа «Гордон» (НТВ) (2003)]; *А* ↓*кро́ме гипо́тезы о Бо́ге/каку́ю ещё гипо́тезу приво́дят проти́вники э́той тео́рии?* [Возникновение биологической информации. Программа «Гордон» (НТВ) (2003)]; *Так во́т/*↓*кро́ме Сату́рна/*

повторя́ю/у все́х плане́т/у Юпи́тера/у Ура́на/у Нептуна́ — то́же е́сть своя́ систе́ма коле́ц [Солнечная система. Программа «Гордон» (НТВ) (2003)] и другие примеры на *кроме* ‘добавления’.

1.5. *Вообще* в разных значениях

Частица *вообще*, описанная в некоторых своих значениях в работе [Апресян 1995], также обладает разными акцентными свойствами в зависимости от значения. В работе [Богуславская 2014] приводится полный список значений *вообще* и отмечается, что в большинстве значений *вообще* допускает или требует главного фразового ударения. Те значения, в которых *вообще 1* не допускает фразового ударения, можно назвать уступительными: *Летом тут вообще бывает тепло, но это лето какое-то холодное*; *Я вообще с ним знакома, но мы давно не виделись.* Как видно из примеров, в обоих случаях *вообще* вводит первую, менее важную ситуацию, которая опровергается во второй клаузе. *Вообще 1* входит в тему, а фокус внимания, прагматическая сила и акцентное выделение падает на вторую часть сообщения, и *вообще* оказывается в безударной позиции. В этом смысле частица *вообще 1* в уступительных значениях устроена так же, как и *только,* и *правда*: уступительные единицы, вводящие менее важную ситуацию, не могут фокусировать на себе внимание — оно фокусируется на более важной ситуации. Ср. пример из МУРКО с безударным уступительным *вообще*: *— Ну что́ тут у ва́с? — Вообще́/ко́мната ↓хоро́шая/све́тлая/но вот ви́дите…* [Родион Нахапетов и др. Не стреляйте в белых лебедей, к/ф (1980)].

Те значения, в которых *вообще* допускает акцентное выделение или требует его, можно обобщить следующим образом: в них *вообще 2* вводит вторую ситуацию, которая каким-то образом противопоставляется первой, вводимой в начальной клаузе [ср. описание в Богуславская 2014]. Соответственно, и в смысле линейного порядка, и в смысле важности ситуация, вводимая *вообще 2,* оказывается в привилегированной позиции фокуса предложения и соответственно акцентуируется. Возможные виды противопоставления и интонационное выделение в разных значениях *вообще 2* (по материалам корпуса МУРКО):

1. *Он халтурщик и ↓вообще человек ненадежный* (вторая ситуация противопоставляется первой как более общая и, соответственно, более «сильная»); *И во́т э́тот во́т виногра́довский повествова́тель-расска́зчик/при то́м/что статья́ про «Пи́ковую да́му» — э́то кла́ссика/и ↓вообще́ замеча́тельное соверше́нно сочине́ние* [Е. В. Падучева. Доклад на конференции «Диалог 2013» (2013)]

2. *— Ну может, ты с ним хотя бы по телефону поговоришь? — Я с ним ↓вообще не буду разговаривать* (вторая ситуация противопоставляется первой как ее тотальное отрицание и, соответственно, как более «сильная»); *Пото́м ко́рпус позволя́ет лу́чше ээ вы́явить структу́ру/то́ есть те́ ле́ммы/кото́рые перепи́сываются из словаря́ в слова́рь/ча́сто ↓вообще́…*

↓вообще́ непра́вильные [Д. О. Добровольский. Доклад на конференции «Диалог 2013» (2013)]; *Ме́жду те́м жёсткие эгоце́нтрики/они́ ли́бо ↓вообще́ не употребля́ются в како́м-то та́м гипотакси́ческом или наррати́вном конте́ксте/конте́ксте/традицио́нного наррати́ва/либо меня́ют сема́нтику* [Е. В. Падучева. Доклад на конференции «Диалог 2013» (2013)].

Противопоставление более общего как более сильного менее общему как более слабому настолько прагматически «заметно», что позволяет вообще 2 нести на себе акцентное выделение даже в тех значениях, где эта частица вводит первую, т. е. прагматически непривилегированную ситуацию: *Я буду говорить о языках ↓вообще и о древнегреческом в частности.*

Итак, семантика **противопоставления**, развивающаяся у языковой единицы, способствует развитию у нее определенных акцентных и коммуникативных свойств — а именно, способности нести на себе главное фразовое ударение и выступать в качестве ремы.

2. Добавление

Следующая важная семантическая идея, которая способствует акцентной и коммуникативной самостоятельности языковой единицы — это идея **добавления**. С этой точки зрения интересно различие значений *кроме 1* и *кроме 2*, обсуждавшееся выше. *Кроме 1* в значении исключения не принимает фразовое ударение, в то время как в значении добавления это становится возможным: *Есть кто-то, ↓кроме Васи, кто решил эту задачу?* В значении добавления в пресуппозиции оказывается идея о том, что имеется некоторый объект, обладающий релевантным свойством, и в ассерцию (и рему) обычно попадает указание на другие такие объекты: *Кроме Васи, задачу решили Петя и ↓Коля.* В ситуации вопроса, когда неизвестно, есть ли другие такие объекты, в фокус внимания попадает сама идея добавления, которая окрашивается отрицанием и противопоставлением и получает способность к акцентному выделению: *Есть кто-то, ↓кроме Васи, кто решил эту задачу?* ≅ ‘Есть кто-то, кто **не является** Васей, кто решил эту задачу’?

Интересно, что и среди значений частицы *еще*, только то из них, которое окрашено семантикой добавления, развивает способность нести на себе фразовое ударение. В работе [В. Апресян 2015] для *еще* предлагается следующий набор значений:

еще 1.1 ‘вдобавок к объектам того же типа’: *Налей мне еще чаю.*

еще 1.2 ‘вдобавок к объектам другого типа’: *Кроме трех газет, он выписывает еще два научных журнала.*

еще 1.3 ‘вдобавок к имеющейся плохой ситуации, усугубляя ее’: *Он опоздал, ничего не сделал и еще хамит?*

еще 2.1 ‘ситуация продолжается’: *Пока она еще работает в старом месте.*

еще 2.2 ‘ситуация наступила раньше ожидаемого’:
Об этом писал еще Аристотель.

еще 2.3 ‘ситуация наступит’: *Он еще пожалеет об этом*.

еще 3 ‘ситуация могла бы быть хуже’: *И это еще цветочки!*

еще 4 ‘говорящий недоволен’: *Что это еще за явление?*

Акцентное выделение возможно только для первого значения, где речь идет о добавлении объектов того же типа: *Он купил ↓еще немного хлеба* (вдобавок к тому, который был). Как и в случае с *кроме 2*, у *еще 1.1.* есть пресуппозиция (а именно, уже имеется некоторое количество объектов с определенным свойством). Ассерцией является указание на добавленные объекты. Это указание вводится именно частицей *еще 1.1* (в ее сферу действия попадает именная группа, обозначающая добавленный объект). Таким образом, *еще 1.1.* находится в ассертивной и, соответственно, рематической части высказывания, что и позволяет этой частице нести акцентное выделение и превращаться в самостоятельную рему. Итак, значение добавления также легко рематизируется, поскольку, по-видимому, содержит некоторую идею противопоставления: речь идет не о тех объектах, о которых уже известно из пресуппозиции, а о других, такого же типа, но не идентичных. Акцентное выделение для *еще 1.1* не обязательно, но возможно; ср. пример такого выделения из МУРКО: *Во́т/и ↓ещё полу́чим сто́лько же* [Жорес Алферов.. Полупроводниковая революция. Наука и общество. Проект Academia (ГТРК Культура) (2010)].

Вообще, по-видимому, наличие пресуппозиции и ассерции в семантической структуре слова способствует развитию у него способности нести акцентное выделение и формировать самостоятельную рему. В качестве еще одного подтверждающего примера можно привести два значения вокабулы *найти*, описанные в работе [Апресян 1995]. *Найти 1* указывает на контролируемое и целенаправленное действие, направленное на поиск потерянного объекта: *Он наконец-то нашел свои ключи. Найти 2* указывает на неконтролируемое и нецеленаправленное действие: *Он неожиданно нашел на улице пять рублей.* Лексема *найти 1* содержит пресуппозицию, что объект был потерян и субъект искал его. Лексема *найти 2* такой пресуппозиции не содержит. Для *найти 1* возможно акцентное выделение и рематизация, для *найти 2* — нет: *Он ↓нашел свои ключи*; **Он неожиданно ↓нашел на улице пять рублей.* Дело в том, что имеющаяся в *найти 1* пресуппозиция в принципе задает два возможных исхода ситуации: объект будет найден и объект не будет найден. *Найти 1* указывает на один из двух изначально возможных исходов; всегда, когда есть две возможности, одна из них может противопоставляться другой. При рематизации *найти 1* подчеркивается идея того, что из двух возможных исходов реализовался положительный, а не отрицательный. Таким образом, как и в случае *еще, кроме, вообще* и других, описанных выше, возникает противопоставление между пресуппозицией и ассерцией, которое привлекает внимание и создает возможность акцентного выделения и рематизации.

3. Прагматические свойства уступительных единиц

Интересные результаты дает не только сопоставление разных значений одного слова, но и сопоставление прагматических свойств семантически близких языковых единиц. В этом смысле богатый материал дает семантическое поле уступительности. Основной уступительный союз, *хотя 1*, не допускает акцентного выделения и не способен к рематизации. Это связано, по-видимому, со следующим. *Хотя 1* имеет следующее значение [В. Апресян 2006]:

(6) *Хотя P, Q* (*Хотя он был болен* [P]*, он пошел на работу* [Q]) = ‘имеет место P; имеет место Q; говорящий считает, что если имеет место ситуация типа P, то обычно или естественно, чтобы не имела места ситуация типа Q.

Как видно из толкования, имеются две ситуации — «проигравшая» ситуация P и «победившая» ситуация Q, и *хотя 1* вводит указание на «проигравшую», более слабую ситуацию. В этом смысле союз *хотя 1* подобен уступительным единицам *только 2, правда 2* и *вообще 1*, которые также вводят указание на менее важную, более слабую ситуацию, и не обладают способностью нести акцентное выделение и рематизироваться. *Хотя 1* обычно входит в тему высказывания. При этом союзы в принципе способны фокусировать на себе внимание и нести акцентное выделение; ср. причинный союз *потому что*:

(7) *В большинстве стран эти конфликты решаются не на основе норм права, а на основе корпоративной этики. То есть взятки там брать нехорошо* ***не*** ***↓потому****, что наступает юридическая ответственность, а* ***↓потому****, что это неприлично* («Еженедельный журнал», 2003.04.01).

Аналогичные фразы с *хотя 1* невозможны, как нам кажется, потому, что *хотя* синтаксически вводит указание на «слабую», «проигравшую ситуацию»:

(8) **Он берет взятки не ↓хотя наступает юридическая ответственность, а хотя это неприлично.*

Ожидаемым образом, так называемые противительные частицы, которые являются семантическими конверсивами к *хотя 1*, т. е. описывают ту же ситуацию, но с другой точки зрения, обладают противоположными прагматическими свойствами. Слова типа *тем не менее, все равно, все-таки, все же* вводят указание не на «проигравшую» ситуацию, а на «победившую»: *И хотя я был лёгкий, тем не менее я ухитрился разбить все стаканы, почти все тарелки и глиняный рукомойник* (М. Зощенко); *Его не звали, и все-таки он пришел*. Это прагматически «сильные» лексемы, они входят в ассертивную часть высказывания, легко принимают на себя акцентное выделение и рематизируются:

(9) *Думаю, ему было чем заняться, кроме нашей семейки. И тем не менее он был здесь* (В. Белоусова).

(10) — *Ведь в нём самом нет ни храбрости и силы льва, ни скорости зайца, ни мудрости филина! Даже зоркость свою он отдаст орлу. Ничего же не останется! — А я всё равно его люблю! — настаивала прекрасная принцесса* (С. Седов).

Для тем не менее позиция обычной темы в принципе невозможна:

(11) **Тем не менее она пришла, хотя ее не звали.*

Если частица *тем не менее* располагается в начале высказывания, то она непременно вводит полемику с предыдущим высказыванием и несет на себе акцентное выделение. С коммуникативной точки зрения, она представляет собой контрастную тему; ср. пример из МУРКО: *Зна́чит не вся́кую ситуа́цию мо́жно прерва́ть/сказа́в «посто́й». Нельзя́ сказа́ть «посто́й»/челове́ку/кото́рый спи́т/и прерва́ть его́ со́н. Нельзя́ сказа́ть «посто́й» челове́ку/кото́рый ду́мает и прерва́ть его́ размышле́ния. Нельзя́ сказа́ть «посто́й» чеку/кото́рый мечта́ет/и прерва́ть его́ мечты́. ↓Те́м не ↓ме́нее у «посто́й» е́сть широ́кий кла́сс ситуа́ций/кото́рый-таки о́н мо́жет прерва́ть* [Е. В. Рахилина. Доклад на конференции «Диалог 2013» (2013).

Таким образом, изменение семантического фокуса и сдвиг его в сторону привилегированной ситуации способствует изменению прагматических свойств лексем на противоположные.

Однако помимо порядка актантов, есть и другие факторы, которые оказывают влияние на способность лексем нести акцентное выделение и рематизироваться. Не все уступительные (т. е. вводящие указание на «проигравшую» ситуацию) лексемы не способны фокусировать на себе внимание. Близкие синонимы союза *хотя*, предлоги *несмотря на* и *невзирая на* и производные от них союзы, а также фразема *несмотря ни на что* способны нести на себе акцентное выделение, а для фраземы *несмотря ни на что* оно является обязательным. Ср.:

(12) *Когда она догадалась, что, ↓несмотря на все её ухищрения с водкой, я её могу бросить, она ушла к нему* (Ф. Искандер).

(13) *Любой, нарушивший эти правила в дальнейшем*, будет выставлен за дверь, *↓невзирая на законы гостеприимства!* (М. Петросян).

(14) *Он заново почувствовал минуты, проведённые с нею, и его охватила жгучая радость, несмотря ни на ↓что* (И. Муравьева).

С чем связана способность единиц *несмотря на, невзирая на* и *несмотря ни на что* нести акцентное выделение и фокусировать на себе внимание? Представляется, что их главное семантическое отличие от союза *хотя 1* — большая степень противоречия между ситуацией P и ситуацией Q: они все указывают на то, что ситуация P была важным и существенным препятствием на пути Q, что не обязательно для союза *хотя 1*. Эта их особенность обусловлена наличием отрицания во внутренней форме, а у фраземы *несмотря ни на что* — еще

и квантора всеобщности. Как отрицание, так и квантор всеобщности — прагматически «заметные», привлекающие внимание элементы; их наличие способствует развитию у *несмотря на, невзирая на* и *несмотря ни на что* нетипичных для уступительных единиц прагматических свойств.

Другой тип уступительных единиц, которые отличаются от *хотя 1* прагматическими свойствами и способностью к акцентному выделению — это путативные единицы *пусть, пусть даже, даже если, хоть бы и*, содержащие указание на **очень высокую степень** ситуации P. В отличие от *хотя 1*, они не фактивны, т. е. указывают не на состоявшиеся ситуации P и Q, а на предположительные; ср. невозможность фактивных контекстов:

(15) **Я пошла пешком, хотя там было десять километров.*

(16) **Я пошла пешком, хоть бы <даже если> там было десять километров.*

В отличие от *хотя 1*, они невозможны в контекстах малой степени:

(17) *Я все успела, хотя он мне немного мешал.*

(18) ??*Я все успею, даже если он будет мне немного мешать.*

Для них предлагается следующее толкование [В. Апресян 2006]:

(19) *Хоть бы и P, Q* = ‘может иметь место P в очень высокой степени; говорящий считает, что если имеет место ситуация типа P, то обычно или естественно, чтобы не имела места ситуация типа не-Q; говорящий уверен, что имеет или будет иметь место Q’ .

Для них возможно акцентное выделение:

(20) *Наверно, человеку свойственно всегда надеяться. Даже если оснований для надежды нет никаких* (В. Быков).

(21) *Великий талант? Да хоть бы и лорд Байрон! Дурные поступки всегда дурные поступки* (Ю. Давыдов).

(22) *Нечего и говорить, что все желали зарегистрировать-таки договор купли-продажи на квартиру номер 37 в доме 4 по улице Новоткацкой — пусть даже и через суд* (А. Волос).

Их прагматические, акцентные и коммуникативные свойства также являются производными от их семантики: они содержат указание на высокую степень — привлекающий к себе внимание семантический компонент, а также на высокую степень уверенности говорящего в том, что нечто будет иметь место. Если фактивность, указание на существование, не окрашенное высокой

степенью (как у *хотя 1*) тяготеет к пресуппозиции и, следовательно, теме, путативность, указание на убежденность говорящего, тем более подкрепленное указанием на высокую степень, тяготеет к ассерции и, следовательно, к реме.

Заключение

Итак, было показано, что акцентные и коммуникативные свойства лексем коррелируют с их семантическими свойствами. В частности, на примере сравнения различных значений слов, а также сравнения семантически близких языковых единиц, была продемонстрирована связь способности нести на себе акцентное выделение и формировать самостоятельную контрастную коммуникативную единицу (самостоятельную рему или контрастную тему) с наличием в значении семантических компонентов противопоставления, добавления и высокой степени. С другой стороны, наличие в значении уступительного компонента, как правило, ограничивает способность лексемы к акцентному выделению и коммуникативной самостоятельности.

Литература

1. *Апресян 1995 — Ю. Д. Апресян.* Избранные труды. Том 1. Лексическая семантика. Синонимические средства языка. М., 1995.
2. *В. Апресян 2006 — В. Ю. Апресян.* Уступительность в языке // Языковая картина мира и системная лексикография / Под ред. Ю. Д. Апресяна. М., 2006. С. 615–712.
3. *В. Апресян 2015 — В. Ю. Апресян.* Словарная статья ЕЩЕ // Активный словарь русского языка. Отв. ред. акад. Ю. Д. Апресян (в печати).
4. *Баранов, Плунгян, Рахилина 1993 — А. Н. Баранов, В. А. Плунгян, Е. В. Рахилина.* Путеводитель по дискурсивным словам русского языка / Научный руководитель проекта Д. Пайар. М., 1993.
5. *Богуславская 2014 — О. Ю. Богуславская.* Словарная статья ВООБЩЕ // Активный словарь русского языка. Тт. 1–2: А–Г. Отв. ред. акад. Ю. Д. Апресян. М.: «Языки славянских культур». 2014.
6. *Богуславский 1985 — И. М. Богуславский.* Исследования по синтаксической семантике. М.: 1985.
7. *Булыгина, Шмелев 1997 — Т. В. Булыгина, А. Д. Шмелев.* Языковая концептуализация мира на материале русской грамматики. М., 1997.
8. *Николаева 1985 — Т. М. Николаева.* Функции частиц в высказывании (на материале славянских языков). М., 1985.
9. *Николаева 2000 — Т. М. Николаева.* От звука к тексту. М., 2000.
10. *Partee 1993 — B. Partee.* On the “Scope of Negation” and Polarity Sensitivity. // E. Hajicova. Ed. Functional Approaches to Language Description. Proceedings of a Conference in Prague, November 24–27 1992. Prague *1993*.

References

1. *Apresjan Yu. D.* Selected works. Volume 1. Lexical semantics [Leksicheskaya semantika]. Moscow, 1995.
2. *Apresjan V. Yu.* (2006), Concession in language [Ustupitel'nost' v yazyke], Linguistic worldview and systematic lexicography [Yazykovaya kartina mira i sistemnaya leskikografiya]. Ju.D. Apresjan, ed. Moscow, 2006. Pp. 615–712.
3. *Apresjan V. Yu.* (2015), Lexicographic entry MORE [Leksikograficheskaja statja ESHCHE], Active Dictionary of Russian [Aktivnyy slovar' russkogo yazyka]. Yu.D. Apresjan, ed. (in print).
4. *Baranov A. N., Plungian V. A., Rakhilina E. V.* A guide to Russian discourse words [Putevoditel' po diskursivnym slovam russkogo yazyka]. Moscow, 1993.
5. *Boguslavskaja O. Yu.* (2014), Lexicographic entry IN PRINCIPLE [Leksikograficheskaya statya VOOBSHCHE], Active Dictionary of Russian [Aktivnyy slovar' russkogo yazyka]. Yu.D. Apresjan, ed. Moscow, 2014. Pp. 245–246.
6. *Bulygina T. V., Shmelev A. D.* Linguistic conceptualization of the world (as exemplified by Russian grammar]. [Yazykovaya kontseptualizatsiya mira (na materiale russkoy grammatiki). Moscow, 1997.
7. *Nikolaeva T. M.* The functions of particles in an utterance (as exemplified by Slavic languages) [Funktsii chastits v vyskazyvanii (na materiale slavyanskikh yazykov]. Moscow, 1985.
8. *Nikolaeva T. M.* From sound to text [ot zvuka k tesktu]. Moscow, 2000.
9. *B. Partee.* On the "Scope of Negation" and Polarity Sensitivity. // E. Hajicova. Ed. Functional Approaches to Language Description. Proceedings of a Conference in Prague, November 24–27 1992. Prague 1993.

СПРАВЕДЛИВОСТЬ VERSUS НЕСПРАВЕДЛИВОСТЬ: МЕТАФОРИЧЕСКИЕ ОСМЫСЛЕНИЯ В СОВРЕМЕННОМ РОССИЙСКОМ ДИСКУРСЕ (ПО МАТЕРИАЛАМ ЦЕНТРАЛЬНЫХ ПЕЧАТНЫХ ИЗДАНИЙ)

Баранов А. Н. (Baranov_anatoly@hotmail.com)

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

JUSTICE VERSUS INJUSTICE: METAPHORICAL INTERPRETATIONS IN MODERN RUSSIAN DISCOURSE (THROUGH TEXTCORPUS OF PRINT MEDIA)

Baranov A. N. (Baranov_anatoly@hotmail.com)

Institute of Russian Language RAS (Vinogradov's intstitute), Moscow, Russia

In the paper words *spravedlivost'* (justice) and *nespravedlivost'* (injustice) in Russian and their corresponding concepts are considered. It is shown that formally words *spravedlivost'* and *nespravedlivost'* are antonyms, because morphologically they differ only in morpheme *ne-* ("no"). But their meanings differ in a more complicated way. Word *spravedlivost'* has an abstract meaning, it denotes a value category. At the meantime extensional set of the word *nespravedlivost'* is another one: it is used for denoting of wide range of situations where features of justice as a value concept are violated. For this reason words *spravedlivost'* de facto is singularia tantum: it has not plural. At the same time the word *nespravedlivost'* (injustice) has in Russian speech bide forms: singular, as well as plural.

Differences in semantics between two words under consideration become apparent in metaphorical models which are used by speakers in interpretation of justice an injustice in Russian public discourse, model of which is text corpus of print media.

Key words: justice, metaphorical model, value category, cognitive approach, descriptor theory of metaphors, metaphorical interpretation

1. Теоретические основания исследования и характер корпусного эксперимента

В ряде исследований слова *справедливость* было показано, что для исследования его семантики и связанной с ним понятийной категории «справедливость»[1] необходимо привлечение дополнительной информации о его функционировании в дискурсе [Baranov 1998], а также о парадигматических связях с другими аналогичными категориями — в частности, со словом *равенство* и соответствующим концептом [Баранов 1990; Baranov 1994]. В данной работе речь пойдет о метафорических осмыслениях концепта «справедливость» в текстах современных российских газет и журналов, а также о тех следствиях, которые влекут выявленные осмысления для семантики слова *справедливость* и связанного с ним концепта.

Для классификации и сбора данных об использованных метафорических моделях привлекался аппарат дескрипторной теории метафоры, в которой контекст употребления метафоры описывается как множество кортежей сигнификативных и денотативных дескрипторов, представляющих, соответственно, область источника и область цели метафорической проекции [Баранов 2014]. Так, метафора *война законов* представляется в дескрипторной теории метафоры в виде двухэлементного множества следующего типа: {<война>, <законодательская деятельность; законодательство>}. Первый дескриптор — «война» — является сигнификативным, а вторые два — «законодательская деятельность», «законодательство» — денотативными.

В качестве источника контекстов метафорического осмысления справедливости использовался корпус Современной русской публицистики, объемом 29 млн словоупотреблений[2]. Всего было выявлено 1550 контекстов употребления слова *справедливость*, причем метафорических осмыслений оказалось 344. Контексты метафорического осмысления были кодированы на языке сигнификативных дескрипторов и введены в компьютерную базу данных, что дало возможность провести анализ частоты употребления метафорических моделей в дискурсе.

Кроме того, анализ справедливости был дополнен аналогичным исследованием слова *несправедливость* и соответствующего концепта. Всего было выявлено 268 контекстов употребления слова *несправедливость*, причем 71 контекст оказался метафорическим осмыслением феномена «несправедливости».

2. Справедливость versus несправедливость

По имеющимся словарным источникам значения слова *справедливость* никак не зависят от грамматических характеристик его словоформ [МАС], [Шведова 2007]. Более того, не налагается никаких ограничений на реализацию

---

[1] Здесь и далее когнитивные категории, концепты, понятия маркируются «елочками» или не выделяются никак, если это понятно из контекста. Слова естественного языка и примеры даются курсивом.

[2] Описание основных особенностей корпуса см. в [Баранов, Добровольский, Киселева, Козеренко 2007].

грамматических категорий. Между тем, в полученной из корпуса выборке на слово *справедливость* содержится 344 контекста использования метафор, причем ни в одном из контекстов не реализуется множественное число. Ср., например, некоторые характерные примеры: *На этот раз торжества* ***справедливости*** *придется подождать. [Итоги]; Конечно, победа* ***справедливости*** *всегда радость. Но она — с таким привкусом горечи! [Известия]*. Аналогичная ситуация обнаруживается и для контекстов, в которых отсутствуют метафорические осмысления. Тем самым, множественное число для слова *справедливость* нехарактерно. Следовательно, лексема *справедливость*, как и большинство слов с абстрактным значением, принадлежит к формам singularia tantum.

Лексема *несправедливость*, непосредственно связанная со словом *справедливость*, формально выглядящая как ее антоним, существенно отличается от слова *справедливость* как по грамматике, так и по значению. Последнее вовсе не антонимично *справедливости*, как это можно было бы предположить по данным словарей. Действительно, первое значение слова *несправедливость* отсылает к семантике прилагательного *несправедливый* [МАС]. В свою очередь, первое значение прилагательного *несправедливый* отсылает к справедливости с отрицанием: ‘поступающий вопреки справедливости, нарушающий справедливость’ [МАС], ‘лишённый чувства справедливости, противоречащий справедливости’ [Шведова 2007]. Таким образом, получается, что *несправедливость* — это антоним к слову *справедливость*.

Исследование контекстов употребления слов *справедливость* и *несправедливость* показывает, что в реальном дискурсе они не антонимы. Так, в отличие от формы *справедливость*, слово *несправедливость* широко используется во множественном числе. Всего в корпусе Современной русской публицистики выявлено 268 контекстов употребления слова *несправедливость*, причем 39 контекстов — употребления во множественном числе (*список несправедливостей; начала несправедливостей — драм и трагедий; ряд несправедливостей*).

Анализ контекстов употребления рассматриваемых лексем показывает, что слово *справедливость* в основном используется в значении ценностной категории (‘свойства, характеристики ситуации, соответствующие представлениям об истине и законе — точнее, правде’ [МАС]). Именно поэтому множественное число данного слова неупотребительно. Для слова *несправедливость* образование множественного числа регулярно, поскольку оно используется в контекстах другой семантики: оно указывает на ситуации нарушения тех или иных норм, правил, представлений о до́лжном, истинном, правильном и т. п.:

> *Руководители Союза приводят примеры всяких* ***несправедливостей****: вот закрыли «Неделю» — 36 человек оказались на улице.* [Московские новости]*;*

> *Имущественный водораздел у очень многих способных, но небогатых людей из провинции вообще отсекает возможность приехать в столицу на вступительные экзамены. А ведь по окраинам талантливых ребят нисколько не меньше, чем в Москве и Петербурге.*

Почему бы не объединить усилия с факультетом, чтобы хоть как-то противостоять этой ***несправедливости****?* [Новая газета]

Есть у меня странное ощущение ***несправедливости*** *природы: папа прожил 55 лет, Толя 62 года. Почти 8 лет разницы.* [Огонек]

В приведенных примерах несправедливость конкретна: закрытие газеты «Неделя» и увольнение сотрудников, сложность поступления в университеты школьников из провинции, уход из жизни людей в разном возрасте и пр. Иными словами, слово *справедливость* относится к ценностным категориям — оно более «идеологично», а *несправедливость* — более конкретно, как правило, референтно и относится к конкретным ситуациям нарушения тех или иных правил, законов, установлений, моральных норм и пр. Именно поэтому слова *справедливость* и *несправедливость* **не антонимы**, а лежащие за ними понятия — концепты — не противопоставлены друг другу — они иные, причем «инаковость» не формирует контрарную оппозицию в духе «A versus не A».

Асимметричность семантики слов *справедливость* и *несправедливость* и лежащих за ними понятий проявляется и на уровне метафорики — метафорических осмыслений, выявленных в представительном корпусе текстов СМИ.

3. Справедливость: метафорические модели и профилируемые смыслы

Распределение сигнификативных дескрипторов по 344 контекстам метафорического осмысления концепта «справедливость» покрывается сравнительно небольшим количеством метафорических моделей. Если ограничить «снизу» частоту дескрипторов пятью вхождениями, то получается следующий набор метафор (сигнификативных дескрипторов):

Таблица 1. Сигнификативные дескрипторы метафор концепта «справедливость» (с частотой 5 и более)

строение/строительство[][3]	148	пункт назначения	16
персонификация[]	134	чел-пациенс[]	12
чел-торжествующий	73	высшее существо	10
объект-предмет[]	28	ресурс	9
чел-агенс[]	24	чел-царствующий	8
пространство()[4]	21	феодализм()	7
движение()	19	чел-требующий	6

[3] Здесь и далее квадратные скобки [] указывают на то, что в данном дескрипторе объединены видовые и родовые дескрипторы. Объединение первых и вторых в данном варианте исследования несущественно.

[4] Здесь и далее круглые скобки () указывают на родовой статус соответствующего дескриптора.

И таблицы 1 видно, что наиболее частотные метафорические модели, используемые для осмысления концепта «справедливость», — это СТРОИТЕЛЬСТВО (вместе с метафорой СТРОЕНИЕ) и ПЕРСОНИФИКАЦИЯ. Высокая частота метафоры СТРОИТЕЛЬСТВА объясняется слабоидиоматичным оборотом (коллокацией) *восстановить справедливость*. 142 из 148 контекстов употребления приходится на примеры использования данного выражения, причем в этих контекстах метафора СТРОИТЕЛЬСТВА относится к стертым метафорам, то есть к устойчивой метафорике, используемой, преимущественно, в фоновом режиме [Баранов 2014]. Широкое использование выражения *восстановить/восстановление справедливости* указывает на то, что ситуация справедливости в наивном языковом сознании рассматривается как нормальное состояние мира.

Контексты использования метафоры ПЕРСОНИФИКАЦИИ, несколько уступающие в количественном отношении метафоре СТРОИТЕЛЬСТВА (134 против 148), существенно более разнообразны. Наибольшее количество приходится на осмысление справедливости как человека, находящегося в психическом состоянии торжества («чел-торжествующий»). Примеры данного типа связаны с фразеологизмом (коллокацией) *торжество справедливости*: ***торжества справедливости*** *придется подождать,* ***справедливость*** *рано или поздно* ***восторжествует.*** В контекстах рассматриваемого типа коммуникативно высвечивается идея реализации ситуации, соответствующей представлениям о справедливости.

Метафора ВЫСШЕГО СУЩЕСТВА в ряде контекстов позволяет провести рефрейминг: приравнять справедливость другому понятию, значимость которого в обсуждаемой ситуации, с точки зрения говорящего, оказывается выше:

> *Он, этот солдат революции, преисполнен уверенности в справедливости,* ***в высшей справедливости возмездия****, которое настигает женщину!* [Огонек]*;*
>
> *Удивительно, но мне нисколько не жаль негодяя, который готовил взрыв на дороге. Когда я замкнул провода и террорист кровавыми кусками взлетел выше деревьев — в этом была* ***высшая справедливость****.* [Московский комсомолец]*;*
>
> *Сталинизм не с неба к нам свалился, это мы с вами растаскивали по домам жалкие пожитки «раскулаченных», если не рукоплескали, то отмалчивались на собраниях, решавших судьбы людей,* ***провозглашали насилие высшей справедливостью****.* [Московские новости].

Сама возможность рефрейминга справедливости указывает на когнитивную важность этой категории в общественном сознании, а также на ее потенциал использования в политической аргументации.

Среди метафор ПЕРСОНИФИКАЦИИ, в рамках которых осмысляется понятие «справедливости», четко противопоставлены агенсные и пациенсные

метафорические проекции. Так, в агенсных проекциях справедливость преследует, наносит удары, требует и т.п.: *от закона уйти можно, а от справедливости — нет; держать удар судьбы и справедливости; справедливость требует признать; историческая справедливость диктует*.

Агенсных осмыслений — 24, а пациенсных — только 11. Типичный контекст реализации пациенсных осмыслений — устойчивое словосочетание *поруганная справедливость*. В пациенсных осмыслениях профилируется смысл несправедливости ситуации. Для профилирования указанного смысла используются и другие метафоры, также относящиеся к пациенсным: ЗАЛОЖНИЦА (*когда справедливость становится ограниченной, тогда она сама превращается в* ***заложницу***), ЖЕРТВА УБИЙСТВА (*У военных прокуроров и судей сейчас задача найти виновных, определить статью, применить закон. Но дай им Бог при этом* ***не убить справедливость***), ЗАЩИЩАЕМЫЙ ЧЕЛОВЕК (*Это просто счастье: быть вместе, ощущать себя великой страной, ощущать себя на хорошей стороне,* ***защитниками*** *морали,* ***справедливости*** *и братства*).

ПРЕДМЕТНАЯ метафора (ОБЪЕКТ-ПРЕДМЕТ) по частоте существенно отличается от ПЕРСОНИФИКАЦИИ и метафоры СТРОИТЕЛЬСТВА: 28 примеров против 148 контекстов метафорической модели СТРОЕНИЯ/СТРОИТЕЛЬСТВА и 134 контекстов использования метафоры ПЕРСОНИФИКАЦИИ. Предметное осмысление существенно более разнообразно в профилировании различных смыслов, чем метафора ПЕРСОНИФИКАЦИИ и СТРОИТЕЛЬСТВА. так, объект-предмет в пространстве может рассматриваться с разных сторон, что позволяет коммуникативно высвечивать различные аспекты ситуации и самого феномена справедливости (*Какая же это справедливость,* ***если она повернута в одну сторону***), а также указывать на сравнительную значимость тех или иных ценностей, осуществляя рефрейминг (*И любовь все так же выше закона. И* ***милосердие выше справедливости***). Предметность дает возможность осмыслять справедливость и как ресурс: *припасти справедливость, кусочек справедливости, справедливость — самый ходовой политический товар*.

Близко к ресурсной семантике осмысление справедливости как ПИЩИ: *хочется то ли Конституции с хреном, то ли* ***севрюжины с социальной справедливостью***; *насытиться справедливостью*.

Внутренняя структура объекта-предмета передает сложность и неоднородность феномена справедливости:

Понятие справедливости ***многослойно,*** *многозначно* [Огонек].

Метафоры ПРОСТРАНСТВА и ДВИЖЕНИЯ в представленном корпусе примеров реализуются почти исключительно как фоновые при осмыслении справедливости как ПУНКТА НАЗНАЧЕНИЯ: *прорыв к справедливости, прийти к справедливости, переход к социальной справедливости, искать пути к согласию и справедливости, объявить курс на социальную справедливость, курс на правовой порядок и социальную справедливость*. Иногда метафоры ПРОСТРАНСТВА и ДВИЖЕНИЯ указывают на неуспешность выбранного способа действия: *тупик социальной справедливости*.

4. Несправедливость: метафорические модели и профилируемые смыслы

Всего для слова *несправедливость* в имеющемся корпусе российских СМИ зафиксирован 71 контекст осмысления в терминах метафор. Состав сигнификативных дескрипторов метафорических моделей, используемых по отношению к понятию «несправедливости», представлен в Таблице 2.

Таблица 2. Сигнификативные дескрипторы метафор концепта «несправедливость» (с частотой 5 и более)

персонификация()	33	ситуация-событие	9
материальная сущность()	22	ограничитель()	8
объект-предмет[]	17	чел-кричащий	7
чел-агенс[]	11	чел-входящий	6
пространство()	10	препятствие	5

Наибольшую частоту имеет метафорическая модель ПЕРСОНИФИКАЦИИ. Треть употреблений этой метафорической модели приходится на метафорические осмысления, в которых «несправедливость» выступает в роли активного действующего лица, которое гонит человека (*именно она, **гонимая болью несправедливостей**, каждый день, как на работу, ходит по всевозможным правоохранительным инстанциям*), делает человека жертвой (***стал жертвой политической несправедливости**; **жертва несправедливости** Патаркацишвили не оставил завещания*), причиняет ему боль (*несправедливость **причиняет мне боль***), задевает человека (*брата сильно **задевала социальная несправедливость***), угнетает (*освободиться от **гнета несправедливости***). Пациенсные осмысления для «несправедливости» отсутствуют.

Второе-третье места делят метафоры МАТЕРИАЛЬНАЯ СУЩНОСТЬ и ОБЪЕКТ-ПРЕДМЕТ (22 и 16 соответственно). Метафора ОБЪЕКТА-ПРЕДМЕТА входит как часть в метафору МАТЕРИАЛЬНАЯ СУЩНОСТЬ. В целом эта метафорическая модель профилирует семантику рационального взаимодействия с феноменом «несправедливости», позволяя его измерять (***большие и маленькие несправедливости** имеют свойство прорастать «гроздьями гнева»*), указывать на возможность его воспроизводства (***конвейер** военной **несправедливости***).

Значительная часть употреблений в рамках метафоры ОБЪЕКТА-ПРЕДМЕТА связана с профилированием смысла ‘ограничения, затруднения, осложнения действий субъекта’: *свинцовое бремя несправедливости; **нести крест несправедливости***; ***столкнуться с несправедливостью***; ***отягощен** преступлениями и **несправедливостями***.

Метафора МАТЕРИАЛЬНОЙ СУЩНОСТИ в рассматриваемом корпусе осмыслений «несправедливости» в целом менее определенна по семантике. Это может быть и профилирование смысла ‘количества’ (*ряд несправедливостей*), ‘измерения’ (*доля несправедливости*), ‘негативных свойств’ (*жгучая несправедливость; ощущаемая несправедливость*), ‘сложности решения проблемы’ (*замкнутое кольцо несправедливости*), ‘значимости’ (*мы оба знаем цену этой несправедливости*).

Дескриптор «ситуация-событие» маркирует в базе данных не столько метафору, сколько родовое наименование некоторого значения: указание на реальное или гипотетическое событие (ситуацию):

> *Руководители Союза приводят* ***примеры всяких несправедливостей****: вот закрыли «Неделю» — 36 человек оказались на улице.* [Московские новости]*;*

> *А все остальные тысяча пятьсот — не напишут. И вот этого одного человека писанину будут где-то там выбирать, это будет двигаться… и так далее. Вы представляете,* ***какая несправедливость происходит****. Ну ладно.* [Огонек]*;*

> *Если вдуматься, то наши критические выступления по* ***конкретным несправедливостям*** *говорят о том же самом.* [Упсальский корпус СМИ]

В приведенных примерах обсуждаются ситуации, которые характеризуются говорящим как несправедливые. Данный тип употребления совершенно нехарактерен для слова *справедливость*.

5. Справедливость versus несправедливость: сравнение метафорических осмыслений

Как было показано выше, слова *справедливость* и *несправедливость* не *антонимы*, а лежащие за ними понятия — концепты — не противопоставлены друг другу. Слово *справедливость* передает семантику ценностной категории: оно более «идеологично», а семантика слова *несправедливость* более конкретна и, как правило, относится к конкретным ситуациям нарушения тех или иных правил, законов, установлений, моральных норм и пр. Легко видеть, что асимметричность семантики слов *справедливость* и *несправедливость* и лежащих за ними понятий проявляется и на уровне метафорики.

Самой частотной метафорой для феномена *справедливости* является метафора СТРОЕНИЯ/СТРОИТЕЛЬСТВА, а для феномена *несправедливости* — метафора ПЕРСОНИФИКАЦИИ. Метафора СТРОЕНИЯ/СТРОИТЕЛЬСТВА реализуется в устойчивом выражении (коллокации) *восстановить справедливость*, которое указывает на то, что **справедливость с точки зрения русского наивного языкового сознания является естественной характеристикой мира.** Действия человека в рамках этой ценности рассматриваются с точки зрения возврата в исходное состояние справедливости или нарушения этого состояния. Таким образом, **феномен несправедливости с токи зрения языкового сознания вторичен.**

Для концепта «несправедливости» самой частотной метафорой оказывается ПЕРСОНИФИКАЦИЯ. значительная часть употреблений приходится на агенсное осмысление этого понятия, причем персонифицированная несправедливость творит что-то плохое: гонит человека, делает его жертвой, причиняет боль, угнетает и т.п. Персонифицированная справедливость (вторая по частоте

метафора при осмыслении справедливости) торжествует в одних ситуациях, требует, диктует — в других, побеждает в третьих и т.д.

Пациенсные характеристики в метафорике справедливости представлены в меньшей степени, хотя справедливость может быть и поруганной, и униженной, она может стать заложницей, жертвой и пр. Таким образом, профилирование свойств метафорами ПЕРСОНИФИКАЦИИ при осмыслении справедливости и несправедливости вполне объясняется значениями соответствующих слов. Исключением можно считать только метафорическую модель ВЫСШЕГО СУЩЕСТВА, позволяющую при метафорическом осмыслении справедливости провести рефрейминг прежнего концепта, изменив его почти до неузнаваемости — в духе оруэлловских упражнений в «новоязе» (ср. *провозгласить насилие высшей справедливостью*).

Наиболее рациональные осмысления феноменов справедливости и несправедливости обеспечиваются метафорой МАТЕРИАЛЬНОЙ СУЩНОСТИ и ее подвида — ОБЪЕКТА-ПРЕДМЕТА. В случае несправедливости эта метафора дает возможность измерять несправедливость (*верх несправедливости*; *большие и маленькие несправедливости, доля несправедливости*), давать ей количественное выражение (*ряд несправедливостей*), указывать на проблемы, которые с нею связаны (БРЕМЯ, НЕСОМЫЙ КРЕСТ, ПРЕПЯТСТВИЕ, ОТЯГОЩАЮЩИЙ ОБЪЕКТ-ПРЕДМЕТ). Довольно разнообразна по профилируемым свойствам в осмыслении феномена справедливости метафора ОБЪЕКТА-ПРЕДМЕТА. Она позволяет рассматривать справедливость как РЕСУРС (*припасать справедливость, справедливость как ходовой товар*), как ПИЩУ (*насытиться справедливостью*), отразить сложность устройства этого концепта (*понятие справедливости многослойно*).

Отметим, что индексы метафоричности обеих рассмотренных категорий очень похожи. Действительно, на 29 млн словоупотреблений корпуса Современной русской публицистики выявлено 1550 контекстов употребления слова *справедливость*, из которых 344 употребления метафорических и 1206 — неметафорических, то есть на одно метафорическое употребление приходится 3,5 неметафорических. Форма *несправедливость* используется в данном корпусе 268 раз, причем 71 контекст является метафорическим осмыслением феномена несправедливости. Тем самым, одному метафорическому контексту соответствует 3,8 неметафорических контекста. Когнитивная теория метафоры утверждает, что чем выше количество метафорических осмыслений понятия, тем оно более важно для общественного сознания. В данном случае и категория справедливости и категория несправедливости имеют очень близкий коэффициент метафоричности: 3,5 и 3,8 соответственно. Таким образом, с когнитивной точки зрения степени значимости и конфликтности этих категорий в общественном сознании (в том виде, в котором оно отражается в СМИ) сходны, если не идентичны.

6. Заключение

Как мы видим, степень метафоричности дискурса о справедливости (и несправедливости) довольно велика. В ранее проведенном исследовании концепта коррупции по аналогичному корпусу СМИ было выявлено 1760

употреблений слова *коррупция*, из которых только 150 употреблений являлись метафорическими [Баранов 2005]. Иными словами, на одно метафорическое употребление пришлось 11,7 неметафорических. **Дискурс о справедливости почти в три раза более метафоричен (3,5 и 3,8), чем дискурс о коррупции. Это указывает на его значимость для общественного сознания**. Данный вывод дополнительно подтверждается исследованием функций ценности «справедливость» в политической аргументации [Баранов 1990].

Довольно странно, что сигнификативное разнообразие обоих рассмотренных концептов невелико. Параметр сигнификативного разнообразия указывает на количество метафорических моделей, используемых при осмыслении исследуемого концепта: «**сигнификативное разнообразие**» целевой области, определяется как отношение числа различных сигнификативных дескрипторов в данной целевой области к числу различных денотативных [Баранов 2014: 140; 466–468]. Поскольку в данном случае концепт (и соответственно денотативный дескриптор) один — справедливость (resp. несправедливость), — то сигнификативное разнообразие определяется количеством метафорических моделей. Если учесть вхождение видовых сигнификативных дескрипторов в родовые (см. таблицу 1 и 2), то для метафорического осмысления справедливости используется семь метафорических моделей (СТРОЕНИЕ/СТРОИТЕЛЬСТВО, ПЕРСОНИФИКАЦИЯ, ОБЪЕКТ-ПРЕДМЕТ, РЕСУРС, ПРОСТРАНСТВО, ДВИЖЕНИЕ, ФЕОДАЛИЗМ), а для несправедливости — четыре (ПЕРСОНИФИКАЦИЯ, МАТЕРИАЛЬНАЯ СУЩНОСТЬ, ПРОСТРАНСТВО, ОГРАНИЧИТЕЛЬ). Объяснение этому можно видеть в том, что большинство выявленных метафорических осмыслений реализуются стертыми (конвенциональными) метафорами. Разумеется, это в определенной степени снижает конфликтность дискурса, которую можно было бы характеризовать по значениям параметра метафоричности как очень высокую. Иными словами, значимость справедливости-несправедливости велика, феномен занимает важное место в ценностной иерархии общества и каждого его члена, но общество в целом притерпелось к реальным ситуациям частого несоответствия идеала актуальному состоянию мира.

Выявленные метафорические осмысления подтверждают анализ семантики справедливости и несправедливости: это не противопоставленные категории, а соответствующие слова не антонимы. При этом справедливость и несправедливость — разные сущности: первая является ценностной категорией, абстрактным понятием, а вторая, скорее, обозначением реальной или виртуальной ситуации, ее свойств.

Литература

1. *Баранов А. Н.* Политическая аргументация и ценностные структуры общественного сознания // Язык и социальное познание, М., 1990.
2. *Баранов А. Н.* О типах сочетаемости метафорических моделей // Вопросы языкознания, 2003, № 2.
3. *Баранов А. Н.* Метафорические модели как дискурсивные практики // Известия АН. Сер. литературы и языка, 2004, том 63, № 1.

4. *Баранов А. Н.* Феномен коррупции в метафорах двух дискурсов // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции Диалог-2005. М., 2005.
5. *Баранов А. Н.* Дескрипторная теория метафоры. М., 2014.
6. *Баранов А. Н., Добровольский Д. О., Киселева К. Л., Козеренко А. Д.* Словарь-тезаурус современной русской идиоматики. М.: Аванта+, 2007.
7. *МАС* — Словарь русского языка в 4-х тт. / под ред. Евгеньевой А. П. М., 1985–1988 [«Малый академический словарь»].
8. *Шведова 2007* — Толковый словарь русского языка с включением сведений о происхождении слов / под ред. Н. Ю. Шведовой. М., 2007.
9. *Baranov A. N.* Justice, Equality and Freedom: The Structure of Value Concepts // P. A. Chilton, M. V. Ilyin, J. L. Mey (eds.) Political Discourse in Transition in Europe 1989–1991, Amsterdam/Philadelphia, 1998, p. 131–145.

References

1. *Baranov A. N.* (1990), Political argumentation and value structures of social consciousness [Politicheskaya argumentatsiya i tsennostnyye sruktury obshchestvennogo soznaniya], Language and social cognition [Yazyk I sotsialnoye poznaniye], Nauka, Moscow.
2. *Baranov A. N.* (2003), About types of metaphor combinations [O tipakh sochetayemosti metaforicheskikh modeley], Questions of linguistics [Voprosy yazykoznaniya], № 2, pp. 73–94.
3. *Baranov A. N.* (2004), Metaphorical models as discursive practices [Metaforicheskiye modeli kak diskursivnyye praktiki], Proceeding of RAS. Series of literature and language [Izvestiya Akademii Nauk, seriya literatury i yazyka], Vol. 63, № 1.
4. *Baranov A. N.* (2005), Phenomenon of corruption in metaphors of two discourses [Fenomen korruptsii v metaforah dvukh diskursov], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2005" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2005"], Bekasovo, pp. 56–62.
5. *Baranov A. N.* (2014), Descriptor theory of metaphor [Deskriptornaya teoriya metafory]. Languages of Slavic cultures, Moscow.
6. *Baranov A. N., Dobrovol'skiy D.O., Kiseleva K. L., Kozerenko A. D.* (2007), Dictionary-thesaurus of modern Russian idioms [Slovar'-tezaurus sovremennoy russkoy idiomatiki]. Avanta+, Moscow.
7. *MAS* — Dictionary of Russian Language in 4 volumes [Slovar' russkogo jazyka v 4-kh tomakh] (1985–1988), Russian Language, Moscow.
8. Dictionary of Russian Language including information about etymology ed. by N. Yu. Shvedova [Tolkovyy slovar' russkogo yazyka s vklyucheniyem svedeniy o proishozhdenii slov] (2007), Azbukovnik, Moscow.
9. *Baranov A. N.* (1998), Justice, Equality and Freedom: The Structure of Value Concepts, Political Discourse in Transition in Europe 1989–1991, Amsterdam/Philadelphia, pp. 131–145.

АВТОМАТИЧЕСКАЯ ИДЕНТИФИКАЦИЯ ОБЩИХ АРГУМЕНТОВ СОЧИНЕННЫХ ГЛАГОЛОВ

Бердичевский А. (aleksandrs.berdicevskis@uit.no),
Экхофф Х. (hanne. m.eckhoff@uit.no)

Университет Тромсё—Норвежский арктический университет, Тромсё, Норвегия

Ключевые слова: синтаксический корпус, сочинение, синтаксис зависимостей, общие зависимые, общие аргументы, русский

AUTOMATIC IDENTIFICATION OF SHARED ARGUMENTS IN VERBAL COORDINATIONS

Aleksandrs Berdičevskis (aleksandrs.berdicevskis@uit.no),
Hanne Eckhoff (hanne. m.eckhoff@uit.no)

UiT The Arctic University of Norway, Tromsø, Norway

We describe automatic conversion of the SynTagRus dependency treebank of Russian to the PROIEL format (with the ultimate purpose of obtaining a single-format diachronic treebank spanning more than a thousand years), focusing on analysis of shared arguments in verbal coordinations. Whether arguments are shared or private is not marked in the SynTagRus native format, but the PROIEL format indicates sharing by means of secondary dependencies. In order to recover missing information and insert secondary dependencies into the converted SynTagRus, we create a simple guessing algorithm based on four probabilistic features: how likely a given argument type is to be shared; how likely an argument in a given position is to be shared; how likely a given verb is to have a given argument; how likely a given verb is to have a given argument frame. Boosted with a few deterministic rules and trained on a small manually annotated sample (346 sentences), the guesser very successfully inserts shared subjects (F-score 0.97), which results in excellent overall performance (F-score 0.92). Non-subject arguments are shared much more rarely, and for them the results are poorer (0.31 for objects; 0.22 for obliques). We show, however, that there are strong reasons to believe that performance can be increased if a larger training sample is used and the guesser gets to see enough positive examples. Apart from describing a useful practical solution, the paper also provides quantitative data about and offers non-trivial insights into Russian verbal coordination.

Key words: treebank, coordination, dependency syntax, shared dependents, shared modifiers, shared arguments, Russian

0. Introduction

This paper reports on an experiment where we use various cues to identify shared arguments in verbal coordinations. The experiment is a part of the full-scale conversion of the SynTagRus dependency treebank[1] of Contemporary Standard Russian (CSR) into the dependency format used in a large family of treebanks of ancient languages originating in the PROIEL corpus.[2] The PROIEL family of treebanks also includes the Tromsø Old Russian and OCS Treebank (TOROT, nestor.uit.no), which contains approximately 300,000 syntactically annotated word tokens divided between canonical Old Church Slavonic (OCS), Old Russian and Middle Russian. By converting the SynTagRus treebank to the PROIEL format, we add a compatible final stage to the TOROT treebank, in effect obtaining a diachronic treebank spanning more than a thousand years, from Late Common Slavic to CSR.

Although the conversion thus has an obvious practical purpose, it also has interesting theoretical implications. In this paper we focus on coordinations and, more specifically, on the analysis of omitted arguments in verbal coordinations. In this area, the PROIEL annotation contains more information than the SynTagRus annotation does: The PROIEL format uses secondary dependencies to mark shared dependents, while the SynTagRus format does not indicate them.

1. SynTagRus and TOROT/PROIEL

Both PROIEL and SynTagRus are dependency formats, both have close links to a particular syntactic framework. The SynTagRus format (Apresyan et al. 2005) is based on the Meaning–Text model (Mel'čuk 1995), which makes it strongly dependent on lexical semantics and gives it a highly granular argument structure representation based on ranked valencies. In other respects, it is the more traditional of the two formats, in that it only allows primary dependencies and only to a limited extent allows empty nodes. The PROIEL format is inspired by Lexical-Functional Grammar (LFG), and the dependency trees are convertible to LFG F-structures (Haug 2010, Haug et al. 2009). This is the origin of several departures from more traditional dependency formats. Structure sharing is indicated by way of secondary dependencies—for external subjects in control and raising structures, but also to indicate shared arguments and predicate identity. The format also systematically uses empty verb and conjunction nodes to account for ellipsis, gapping and asyndetic coordination. Argument representation is less granular than in SynTagRus, and the labels are largely based on morphosyntactic features: transitive objects (OBJ) are distinguished from oblique objects (OBL). In addition, complement clauses, passive agents, and arguments with

[1] Developed by the Laboratory of Computational Linguistics at the Institute for Information Transmission Problems (to whom we are grateful for the use of the offline version of SynTagRus and advice on its usage), found at http://ruscorpora.ru/search-syntax.html

[2] Developed by the members of the project *Pragmatic Resources in Old Indo-European Languages*, found at foni.uio.no:3000

external subjects, e. g. control infinitives, have separate labels, resp. COMP, AG and XOBJ. The two formats also take different approaches to coordination, as illustrated in Fig. 1 and 2, and further discussed in Section 2.

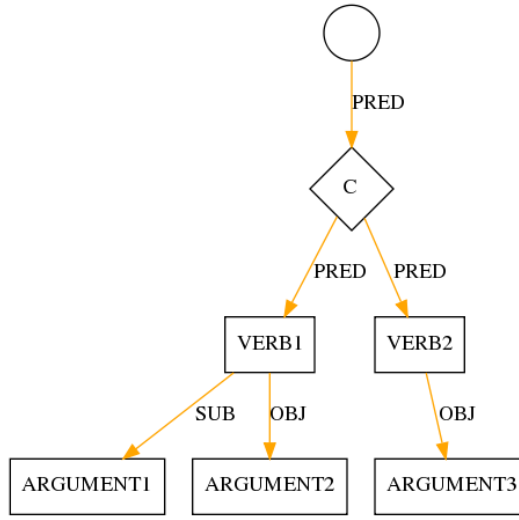


Fig. 1. PROIEL-style coordination. The conjunction is the head of the coordinated nodes, and its outgoing relations are the same as the ingoing one

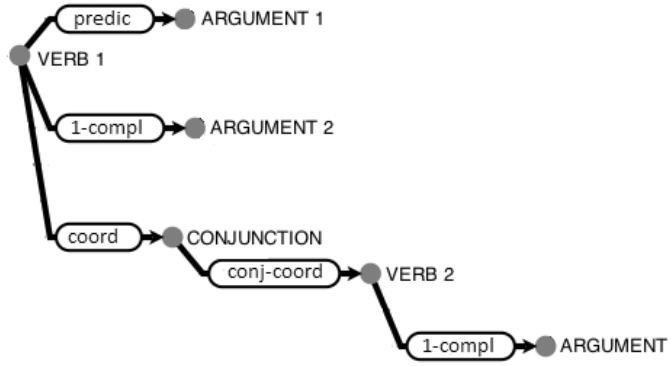


Fig. 2. SynTagRus-style coordination. The first conjunct is the head of the conjunction, which again is the head of the second conjunct, they are connected through a set of special "coordinative" relations

We managed to convert SynTagRus to the PROIEL format with high accuracy using a rule-based algorithm (excluding 528 sentences where the algorithm identified a structure that would most likely not be converted correctly, which leaves us 58,712

sentences). A spot check of 50 random converted sentences (763 words, including empty tokens) gives an unlabeled attachment score (UAS) of 0.96 and a labeled attachment score (LAS) of 0.93. The main issues encountered in the conversion process had to do with coordination and null verb insertion, since the necessary information was not always recoverable in the SynTagRus data. In many other cases, however, it was possible to recover missing information using lexical, morphological and part-of-speech cues. For instance, secondary dependencies indicating external subjects (e. g. of control infinitives and predicative complements) were inserted with 0.96 accuracy. In the current experiment, we insert secondary dependencies indicating shared arguments in coordinations. We are not aware of any highly successful previous attempts at this task. Zeman et al. (2012:2738), converting various treebanks to the Prague style, note that they "apply a few heuristics", but cannot recover shared dependents reliably. Mareček and Kljueva (2009:28–29), converting SynTagRus to the Prague style, note that they apply "a couple of rules", but "by far not all cases can be solved."

2. Coordination: theoretical and practical importance

Dependency grammar is fundamentally based on asymmetric relationships: one node is the head and another node is its dependent, i. e. it is based on subordination. Coordination, therefore, constitutes a problem to all dependency annotation, since it is symmetrical in nature: in paratactic relationships, there is no head, rather, nodes are equal and have the same status. Any dependency format will therefore have to coerce coordinations into hierarchical structures in some way, and it is not obvious what the best choice is. Popel et al. 2013 offer a comprehensive typology of coordination styles, recognising three major approaches: 1) Moscow-style, where the first conjunct is the head, the second conjunct is a dependent on the first, the third is a dependent on the second etc. As regards conjunctions, the policies can vary. In SynTagRus (Fig. 2), they are typically placed between the conjuncts according to the word order: the first conjunction is a dependent on the first conjunct, the second conjunct is a dependent on the first conjunction, etc. 2) Prague-style, where the conjunction is the head, and all conjuncts are its dependents. 3) Stanford-style, where the first conjunct is the head, and the following conjuncts and conjunctions are its dependents. There are different advantages and drawbacks to each solution. The Moscow-style solution has considerable advantages when it comes to simplicity of potential syntactic queries (the first conjunct will be the direct dependent of its true head), whereas the Prague-style solution makes it possible to render complicated stacked structures better, and also allows encoding of shared dependents.

The PROIEL format (Fig. 1) employs a variety of the Prague-style solution, also including the use of null verbs in the case of gapping and other types of verb ellipsis, and null conjunctions in the case of asyndetic coordination. The PROIEL format also uses secondary dependencies to indicate predicate identity and shared dependents. The latter feature is of particular importance for large-scale studies of argument structure, since it makes it possible to extract more complete argument frames. For some possible uses of such data, see Berdičevskis and Eckhoff 2014.

3. Materials and methods

3.1. Objective, key notions and limitations

Our objective is to insert correct shared-argument secondary dependencies in verbal coordinations. We assume that such dependencies can potentially be present when and only when one of the two coordinated verbs has an argument slot filled, while the other verb has the same slot unfilled. The empty-slot criterion is illustrated in Fig. 1: both Verb1 and Verb2 have the object slot filled, and thus none of the objects can be shared. Verb2, however, does not have a subject, while Verb1 does, and thus it is possible that Argument1 is a subject for Verb2, too.

Obviously, coordinated groups can contain more than two verbs, and our samples do include such cases, but at any given point in time we will be looking at two verbs only. More specifically, a *datapoint* in our analysis will be a pair {Argument; Verb}. In Fig. 1, the pair is {Argument1; Verb2}; in an example such as

(1) *Потом он посмотрел на нее и обрадовался*[3]
'Then he looked at her and was pleased'

the available pairs are {он (SUB) 'he'; обрадовался 'was_pleased'} and {на (OBL) 'at'; обрадовался 'was_pleased'}. For every datapoint, we will have to make a decision whether a secondary dependency should be inserted or not (it should in the first of the two pairs mentioned in the previous sentence, but should not in the second). For brevity's sake, we will call drawing a secondary dependency an *adoption* of an argument by a verb (extending the parent-and-children metaphor that is widely used in syntax).

In this experiment, we attempt to identify cases of argument adoption only, i.e. we focus on the relations SUB, OBJ, OBL, XOBJ, COMP, AG (Section 1).[4] In addition, we impose several other limitations on our dataset: we exclude empty verbal nodes[5], participles (since they are expected to display more limited argument frames) and some complex infrequent sharing cases not covered by Fig.1, such as arguments shared by one conjunct verb and a verb in a dependent clause daughter of another conjunct verb.

---

[3] Ju. Kazakov, *Dvoe v dekabre*.

[4] For the relation OBL, the empty-slot criterion is more relaxed: the adopting verb is not required to have the OBL slot unfilled. In other words, since one verb can have more than one outgoing OBL (which is forbidden for the other five relations), verbs are permitted to adopt OBLs even if they already have OBLs as their primary dependents. For the other relations, if there are several candidates that satisfy the empty-slot criterion, only one of them will be selected (the most plausible one, as determined by the algorithm described in Section 3.4).

[5] Elided copulas and cases of verbal gapping

3.2. Data

We used a number of probabilistic features to identify adoption (Section 3.3). For some of the features, we calculated values using the whole converted version of SynTagRus. For some of the features, however, a training set with correct secondary dependencies already inserted was required. For that purpose, we manually annotated a small sample.

At the first stage, the sample included 297 sentences. The sentences were randomly drawn from the whole corpus, the only requirement being that each sentence should contain at least one potential adoption case, i. e. at least one construction satisfying the criteria described in Section 3.1.[6] If sentences contained critical errors that distorted coordination (mostly as a result of conversion, sometimes SynTagRus misannotations), we excluded them, which resulted in 267 annotated sentences.

The analysis of this sample, however, showed that different arguments behave very differently with respect to how often they occur as potential adoptees and with respect to how often they actually are adopted, i. e. to their adoption rate (cf. Table 1). SUBs are frequent and likely to be shared, which gives us a lot of both positive and negative examples. OBJs and OBLs are frequent, but much less likely to be shared, which gives us a lot of negative examples, but very few positive ones. XOBJs are relatively frequent, but very unlikely to be shared, and are thus even more problematic. COMPs and especially AGs are infrequent, and it is not even possible to calculate their expected adoption rate.

To estimate the adoption rate for infrequent arguments, we turned to the TOROT data, which are tagged for secondary dependencies, restricted to the Codex Marianus (OCS), which has been proofread and submitted to comprehensive consistency checks (Table 1, columns 3 and 4). We assume that tendencies in argument sharing are approximately the same in OCS and CSR.

Given the extremely low frequency of positive examples for XOBJ (1), COMP and AG (0) and low estimated adoption rates, we focus mostly on inserting secondary dependencies for the other three arguments. In order to get more positive examples for OBJ and OBL, we extracted another 91 SynTagRus sentences (12 were excluded as containing critical errors), this time requiring that they contain a potential adoption case of one of these relations. The total sample size was 346 sentences, 1103 datapoints (Table 1). The increase in positive OBJ and OBL examples after annotating the extra sample was very small, the total numbers being resp. 16 and 8. Still, we decided to investigate whether it is possible to insert secondary dependencies for these arguments, relying on the sparse data available.

---

[6] For technical reasons, the empty-slot criterion was applied in its strong form to all six relations (cf. footnote 4). This means that some potential cases of OBL adoption might not have made it into the sample, but their number is most probably negligible and could not have affected the results significantly.

Table 1. Number of potential adoption cases and adoption rate (#real adoptions/#potential adoptions) for SynTagRus (estimated on a sample of 346 sentences, see main text) and the Codex Marianus (6350 sentences)

Argument type	Adoption rate (SynTagRus sample)	Potential adoptions (SynTagRus sample)	Adoption rate (Marianus)	Potential adoptions (Marianus)
SUB	0.80	261	0.59	901
OBJ	0.06	251	0.06	895
OBL	0.02	393	0.01	1,409
XOBJ	0.01	158	0.01	472
COMP	0.00	36	0.02	94
AG	0.00	4	0.00	9

3.3. Probabilistic features

By combining our experience in tagging shared dependents manually, theoretical reasoning about argument frames, and the trial-and-error method, we identified four features expected to be the most informative about the adoption status of a given datapoint.

The first feature is the probability of a potential adopter having a potential adoptee as an argument. For the pair {он; обрадовался} in example (1), that would be the probability of the lemma обрадоваться having a subject (as opposed to it being subjectless), which is 0.70. The second one is the probability of a potential adopter having an argument frame that would consist of its own primary arguments *and* a potential adoptee. For example (1), that would be the probability of обрадоваться having a frame V+sub+obl (0.10). These data were calculated using the whole SynTagRus corpus and did not require any manual tagging of secondary dependencies.[7]

It would seem reasonable to expect that for low-frequency verbs these data can be misleading rather than useful. However, when we tried excluding these features for low-frequency potential adopters, that led to decrease in performance. In other words, very little information turns out to be better than no information (see Berdičevskis and Eckhoff 2014:11 for a similar conclusion about argument frames as an information source).

The third feature is the probability of a particular argument being an adoptee, see Table 1, column 1. We tried excluding this feature as disfavouring non-subjects too strongly (or excluding it for non-subjects only), but that led to decrease in performance, both for subjects and non-subjects.

---

[7] Full data on all the verbs and other data not reported in the paper can be found at the TROLLing Dataverse (http://opendata.uit.no/), hdl: 10037.1/10174.

Table 2. Adoption rate for different positions of a potential adoptee, calculated for all arguments together; subjects only; non-subjects only

Position of a potential adoptee	Adoption rate (all arguments)	Adoption rate (subjects only)	Adoption rate (non-subjects only)
FL (on the first conjunct and to the left of it)	0.75	0.93	0.17
FR (on the first conjunct and to the right of it)	0.06	0.65	0.02
M (on a middle conjunct)	0.03	0.21	0.01
LL (on the last conjunct and to the left of it)	0.01	0.06	0.00
LR (on the last conjunct and to the right of it)	0.03	0.25	0.02

The fourth feature is the probability of a potential adoptee being adopted if it is found in a given *position*. We distinguished between five positions: FL, FR, M, LL and LR (Table 2). The subjects, however, again create a strong bias: first, they are the most frequent adoptees, second, they occur most frequently to the left of the verb, third, according to SynTagRus guidelines for annotators, shared subjects have to be placed on the *first* conjunct, while other arguments have to be placed on the nearest conjunct regardless of whether they are shared or private (Leonid Iomdin 2015, p. c.). Given all that, it seemed reasonable to calculate position adoption rate separately for subjects and non-subjects, which we did (Table 2, columns 3 and 4), but using separate scores, again, led to decrease in performance, so we used data from column 2 in our final evaluation.

Values for these two features were calculated using the manually tagged sample described in Section 3.2. We tried including some other features (such as, for instance, the probability of a given conjunction allowing its children to adopt its grandchildren), but that did not lead to any increase in performance.

To sum up, for every datapoint four features were identified, whose values are probabilities between 0 and 1.

3.4. Algorithm

We devised a very simple guessing algorithm. For every datapoint in a training set, it calculates an average of the four probabilities (note the result is not a probability in itself). All datapoints with an average higher than a certain cutoff *c* are considered to be cases of real adoption, all datapoints with the average lower than c are not. The algorithm finds the optimal value of c for the training set (by trial and error, i. e. by trying all possible values of c and selecting the one which gives the highest accuracy for the training set), and then applies the calculated feature values and c to the test set.

We also trained a support vector machine using SVMlight (Joachims 1999) with a radial basis function kernel (other parameters default for SVMlight). SVM provides a slightly higher accuracy than our average-based guesser per se, but when both outputs are corrected using deterministic rules (Section 3.5), our guesser outperforms SVM, especially as regards infrequent arguments types. For this reason, SVM performance is not reported here.

3.5. Deterministic Rules

In addition to the statistical algorithms, we also implemented several deterministic rules. All rules concern only subjects and can only predict negative outcomes (i. e. the absence of adoption) that override statistical guesses. The rules are as follows: a first-person potential adopter cannot adopt subjects that have lemmas different from *я* 'I', *мы* 'we' and *сам* 'self'; a second-person potential adopter cannot adopt subjects that have lemmas different from *ты* 'thou', *вы* 'you' and *сам* 'self', unless it is in the imperative; if a potential adopter has person, number, gender or mood (relevant values of the PROIEL mood category are indicative, imperative and infinitive) different from the potential subject adoptee's real parent, then the adoption is impossible. Thus, we avoid adoptions with obvious person agreement clashes. In addition, there are some rules specific to the PROIEL format.

These rules never fail on our dataset and prevent a small number of false positives.

4. Results and discussion

The guesser's performance was evaluated using 5-fold cross-section validation. Since the classes are highly skewed (adoptions are rare), we report not only accuracy, but also F-score, precision and recall for every argument type (Table 3).

Table 3. Performance of the average-based guesser (with rule-based correction)

	Accuracy	F-score	Precision	Recall	Datapoints	Real adoptions
Overall	0.97	0.92	0.95	0.89	1,100	234
SUB	0.95	0.97	0.97	0.96	261	209
OBJ	0.96	0.31	0.40	0.28	250	16
OBL	0.97	0.21	0.20	0.25	392	8

The overall good results are mostly achieved through excellent performance on SUBs. For OBJs and OBLs, the total number of positive examples is extremely small, and hence the guesser is providing many false negatives. This results in high accuracy, but low F-scores.

Manual error analysis confirms that most false negatives are non-subjects. Typical reasons why they get low average scores are low scores for the "relation"

and "position" features. In example (2), the guesser misses a case of OBJ adoption: *дискредитировать* 'discredit' should adopt *нас* 'us'. The reason is the low adoption rate for OBJs and a low position score for FR.

(2) *Им важнее выдавить нас из страны, дискредитировать, доказать, что мы воруем деньги наблюдателей*[8]
'It is more important to them to squeeze us out of the country, to discredit [us], to prove that we are stealing the observers' money.'

When false negatives are subjects, they are almost always in a non-typical position (postverbal, or depending on a non-first conjunct, or both). For instance, the guesser misses the adoption of the postverbal SUB *она* 'she' of the first conjunct in example (3) by the verb *опустила* 'lowered':

(3) *Ты не сердись,—торопливо сказала она и опустила глаза.*[9]
'Don't be angry, she said hastily, and lowered her eyes.'

As mentioned above, excluding the "position" feature did not lead to increase in performance: even non-subjects tend to be shared more often when they are in FL position (Table 2). A possible reason is that *topical* elements are more likely to be shared.

False positives are less numerous (which is good: for linguistic uses of the corpus, it is better to miss real adoptions than to insert false ones). Interestingly, more than half of the cases identified by the guesser as false positives at the intermediate work stage turned out to be human annotation errors (i. e. cases where we should have inserted secondary dependencies when tagging the extracted sample, but failed to do so). This means that the algorithm can have a practical application as an error identification device in a manually annotated treebank, and an experimental application of the algorithm to the OCS part of TOROT has already confirmed this.

Given the high performance for SUBs, it is reasonable to expect that similar results could be achieved for OBJs and OBLs, if several hundred sentences containing potential adoption cases were annotated and thus at least several dozens positive examples for each relation were collected. With more data, there would also be more possibilities to fine-tune the features to avoid the excessive punishment of non-subject relations.

5. Conclusions

We have described a simple algorithm which allows us to identify shared arguments with high accuracy and F-score. While the performance is excellent on subjects, F-scores are low for other relations. For objects and obliques, the algorithm has a clear

[8] E. Masjuk. "Lilija Shibanova: Vladimir Vladimirovich, vy bol'ny shpionomaniej", *Novaja gazeta,* 45, 2013.

[9] Ju. Kazakov, *Dvoe v dekabre.*

potential of achieving much better results, if more sentences are manually tagged in order to collect more positive examples.

Our solution is a practical contribution, useful both for our specific purposes (the conversion of the SynTagRus to the PROIEL format and subsequent diachronic studies of verbal argument frames) and more general applications (shared subject identification can, for instance, be important for agreement studies). It can potentially contribute to theoretical linguistics, too, by providing quantitative data about some tendencies in Russian coordination. It can, for instance, be tested whether our observation that topical arguments are more likely to be shared holds.

References

1. *Apresyan Yu. D., I. M. Boguslavskij, B. L. Iomdin, L. L. Iomdin, A. V. Sannikov, V. Z. Sannikov, V. G. Sizov, L. L. Tsinman.* (2005), Syntactically and semantically annotated Russian corpus: state of the art and perspectives [Sintaksicheski i semanticheski annotirovannyj korpus russkogo yazyka: sovremennoe sostoyanie i perspektivy], Russian National Corpus [Nacional'nyj korpus russkogo jazyka: 2003–2005]. Indrik, Moscow, pp. 193–214.
2. *Berdičevskis A., H. Eckhoff.* (2014), Verbal constructional profiles: reliability, distinction power and practical applications, Proceedings of the Thirteenth International Workshop on Treebanks and Linguistic Theories (TLT13), Tübingen, pp. 2–13.
3. *Haug D.* (2010), PROIEL guidelines for annotation. http://folk.uio.no/daghaug/syntactic_guidelines.pdf
4. *Haug D., M. Jøhndal, H. Eckhoff, E. Welo, M. Hertzenberg, A. Müth.* (2009), Computational and Linguistic Issues in Designing a Syntactically Annotated Parallel Corpus of Indo-European Languages, Traitement Automatique des Langues, Vol. 50:2, pp. 17–45.
5. *Joachims T.* (1999), Making large-Scale SVM Learning Practical, Advances in Kernel Methods—Support Vector Learning, MIT-Press, Cambridge, pp. 41–56.
6. *Mareček D., N. Kljueva.* (2009), Converting Russian Treebank SynTagRus into Praguian PDT Style, Proceedings of the Workshop on Multilingual Resources, Technologies and Evaluation for Central and Eastern European Languages, pp. 26–31.
7. *Mel'čuk I.A.* (1995), Russian Language in the Meaning-Text Model [Russkij jazyk v modeli "Smysl–Tekst"], Škola "Jazyki russkoj kul'tury", Vienna.
8. *Popel M., D. Mareček, J. Štěpánek, D. Zeman, Z. Žabokrtský.* (2013), Coordination Structures in Dependency Treebanks, Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Sofia, pp. 517–527.
9. *Zeman D., D. Mareček, M. Popel, L. Ramasamy, J. Štěpánek, Z. Žabokrtský, J. Hajič.* (2012), HamleDT: To Parse or Not to Parse?, Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), Istanbul, pp. 2735–2741.

ЗАТРУДНЕНИЯ ПРИ ПОРОЖДЕНИИ СЛОВ В ДИСКУРСЕ И ИХ ФОРМАЛЬНЫЕ МАРКЕРЫ: НОРМА И ПАТОЛОГИЯ, ИЛИ О НЕДИСКРЕТНОСТИ НОРМЫ В ЯЗЫКЕ И РЕЧИ[1]

Бергельсон М. Б.[1], **Акинина Ю. С.**[1], **Драгой О. В.**[1], **Искра Е. В.**[1,2], **Худякова М. В.**[1]

[1]Национальный исследовательский университет Высшая школа экономики, Москва, Россия;
[2]Центр патологии речи и нейрореабилитации, Москва, Россия

Ключевые слова: нарратив, норма, афазия, речевые ошибки, лексические затруднения, фальстарты, самоисправления говорящего, дискурсивные маркеры

MARKERS OF WORD PRODUCTION DIFFICULTIES IN NORMAL AND CLINICAL DISCOURSE PRODUCTION: CONTINUITY OF NORM IN LANGUAGE AND DISCOURSE

Bergelson M. B.[1], **Akinina Yu. S.**[1], **Dragoy O. V.**[1], **Iskra E. V.**[1,2], **Khudyakova M. V.**[1]

[1]National Research University Higher School of Economics, Moscow, Russia;
[2]Center for Speech Pathology and Neurorehabilitation, Moscow, Russia

Aphasia is language impairment due to brain damage. Word-finding and word-retrieval problems can be very prominent in the speech of people with aphasia, being detectable in almost every aphasic speaker. On the other hand, word-finding difficulties and speech errors can sometimes occur in speech of neurologically healthy people. It is assumed that the same psycholinguistic levels of word-retrieval breakdown can account for the mistakes of both groups. In the meanwhile, retrieving of a single word from mental lexicon is not the only possible level of hindrance for a speaker: referential and lexical choices that take place at more general discourse and pragmatic level can also be disturbed.

[1] Настоящее исследование поддержано грантом РФФИ «Мозговые основы порождения дискурса: нарушения нарратива у пациентов с поражениями мозга», 13-06-00614 А

The Russian CLiPS—Russian CLinical Pear Stories—is a corpus of film-elicited narratives retrieved following (Chafe, 1980) methodology from healthy and language-impaired cohorts. The aim of our research was to investigate the characteristics of formal markers of word retrieval difficulties in narratives of neurologically healthy people and people with aphasia. Three types of markers were considered (discourse markers, false starts and self-corrections) in the nominations of common referents of Pear stories narratives. The markers at different breakdown levels are qualitatively analysed, creating a platform for future analysis.

Key words: narrative, norm, aphasia, word production, false starts, self-corrections, discourse markers

1. Введение

1.1. *Корпуса устной, или звучащей, речи* строятся и анализируются с позиции когнитивных процессов, происходящих в режиме реального времени при порождении речи (Кибрик, Подлесская, 2009; Коротаев, Кибрик, 2008).

Однако лингвисты имеют дело с текстами как результатами дискурсивных событий, а это означает, что анализ происходит с позиций интерпретатора, то есть оффлайн, или асинхронно (Федорова, 2014: 104). Перед интерпретатором, в отличие от говорящего, стоят несколько другие задачи. В частности, именно интерпретатор дискурса заинтересован в его максимальной «гладкости» и соответствии норме языка. Норма, грамматическая правильность и гладкость текста — как бы эти термины ни понимались в различных областях лингвистических исследований — представляет собой совокупность ожиданий со стороны интерпретатора дискурса, требующего соблюдения своих интересов. Эти интересы заключаются в балансе между большим количеством элементов кода (Панов 1968), эксплицирующих оттенки смысла, и конвенциональностью этих элементов (имен референтов, дискурсивных маркеров, средств оценки), позволяющей выразить, наоборот, менее специфицированные смыслы.

1.2. Традиционно представление о *норме* резервировано только в социолингвистическом смысле, только в отношении всей языковой системы. Однако при анализе нарративного дискурса возникает определенная проблема. На уровне микроанализа дискурса лингвисты предпочитают использовать оппозицию «правильно ~ неправильно», применяемую к грамматическим и лексическим ошибкам, а также — для устного дискурса — к нарушениям произносительных норм. Но на уровне макроанализа (например, соотношения компонентов, использования типов дискурсивного изложения, средств оценки) отклонения — в ту или иную сторону — рассматриваются как стратегии говорящего, в разной степени обладающего мастерством, искусством, рассказа, что приводит к «успешности» или «не успешности» рассказа. Таким образом норма и правильность оказываются разнесенными по разным, весьма далеким друг от друга зонам внимания исследователей — норма в социолингвистике, правильность в грамматике и отчасти в дискурсе. Кроме того,

помимо противопоставления нормы/не-нормы и правильности/неправильности, в других областях наук о языке, в клинической лингвистике присутствует оппозиция нормы и патологии. На уровне грамматических конструкций и лексических выборов к ней применимы критерии грамматичности и приемлемости (Chapman, Routledge, 2009), и в этом случае оппозиция нормального и не-нормального выглядит более или менее бинарной. Однако при исследовании устного дискурса граница между тем, что считать нормальным и патологичным в ситуации рассказывания, становится зыбкой. Устный дискурс и в норме характеризуется паузами, хезитациями, фальстартами и затруднениями, связанными с выбором подходящей номинации, которые могут восприниматься как избыточные по сравнению с «идеальным», или «нормальным» рассказом. Клинический устный дискурс характеризуется как минимум этими же явлениями, но в гораздо большем масштабе. Тогда в какой степени можно говорить о дискретности нормы и патологии при порождении устного дискурса? Есть ли формальные показатели, отличающие затруднения говорящего при выборе номинаций в дискурсе в случае клинического и обычного дискурсивных взаимодействий?

Цель настоящего исследования заключается в попытке пересечения двух парадигм — клинически-лингвистической и дискурсивной — для исследования затруднений и их вербальных маркеров припорождении устного дискурса у пациентов с афазией в сравнении со здоровыми носителями русского языка.

1.3. *Афазия* — потеря или нарушение языковой способности вследствие поражения головного мозга (Benson & Ardila, 1996). Типы афазий выделяются на основе характерных языковых и сопутствующих смежных когнитивных нарушений, преобладающих в речи пациента, а также анатомического очага поражения. Существуют различные традиции классификации афазий (Ardila, 2010), но трудности с поиском (извлечением, порождением) отдельных слов, как в ситуации тестирования, так и в спонтанной речи, испытывают практически все люди с афазией (Laine, Martin, 2006).

И в то же время затруднения при порождении слова могут быть выявлены и у здоровых носителей языка. Как резюмируют авторы (Laine, Martin, 2006), речевые ошибки включают в себя добавления, удаления, замены или передвижения (предвосхищение, персеверация, обмен) лингвистических единиц — слов, связанных морфем, слогов, фонем и, возможно, даже фонетических признаков. Ошибки в речи людей с афазией в большинстве случаев качественно сходны с теми, что наблюдаются в ненарушенной речи. Отличия проявляются количественно (варьируя в зависимости от грубости поражения), а также в пропорциональном распределении типов ошибок (в зависимости от того, какие этапы порождения слова нарушены и в какой степени). Таким образом, на уровне психолингвистических механизмов извлечения отдельного слова можно говорить об общей функциональной природе речевых сбоев в норме и патологии.

В настоящее время разработано множество моделей того, как происходит извлечение слова из ментального лексикона. Ониконкретизируют его отдельные этапы и механизмы их взаимодействия (Laine, Martin, 2006) и в том или другом виде применимы при анализе данных, полученных на выборках речевой нормы

и патологии, что и должна обеспечивать адекватная модель. (Nickels, 1997). Современные общепринятые в западной психолингвистике модели порождения слова в общем случае постулируют существование трех этапов — извлечение значения (семантика), извлечение формы слова (выходной фонологический лексикон) и формирование релевантной фонологической программы (выходной фонологический буфер) (Laine, Martin, 2006). В нашей работе мы будем пользоваться этой упрощенной, но самодостаточной в рамках настоящего исследования классификацией.

1.4. Следует отметить, что причины затруднений при порождении слова не сводятся только к неправильному срабатыванию механизма извлечения. Здесь также оказываются задействованы явления маршрутизации информации при ее вербализации, формирующие семантическую структуру высказывания на ранних этапах ее порождения. Неудачным или менее удачным с точки зрения развития дискурса может быть референциальный выбор (обозначения референта с использованием полной именной группы, указательного местоимения и т. д.). То же самое верно для выбора между различными номинациями участника описываемой ситуации. Данные процессы подчиняются общим дискурсивно-прагматическим закономерностям, как то — степени полноты знания о внешнем мире, степени точности построения модели текущего сознания адресата. Так, затруднения при выборе номинации могут быть связаны с недостаточностью информации у говорящего, что заставляет его колебаться, например, при выборе между именами *фермер, садовник, сборщик, владелец груши*. Затруднения при осуществлении референциального выбора также не связаны с механизмом извлечения слов из ментального лексикона. Здесь колебания связаны с оценкой говорящим конкретного момента данного дискурсивного события. Колебания в выборе между полной именной группой или указательным местоимением зависят от оценки говорящим текущего состояния сознания адресата и корректировкой этой оценки.

Таким образом, как в нормальной, так и в патологической речи теоретически можно выделить как минимум два потенциальных источника затруднения при использовании лексической единицы: первый — прагматически и дискурсивно мотивированные сбои, связанные с колебаниями в выборе номинации; второй — сложности извлечения единицы ментального лексикона в процессе речепорождения. Это является основанием для постановки задач настоящего исследования: ожидая встретить в речи и той, и другой группы затруднения и ошибки, мы хотим проверить, будет ли принципиально отличаться их словесное оформление в дискурсе, а именно: *фальстарты, самоисправления и вербально выраженные дискурсивные маркеры поиска*. Качественные сходства или, наоборот, принципиальные отличия позволят судить о валидности совмещения дискурсивной и клинико-лингвистической парадигм и дадут основания для постановки дальнейших конкретных исследовательских вопросов.

В качестве материала для анализа мы используем выборку текстов из корпуса «Истории о грушах в клинических популяциях»[2].

---

[2] Мы выражаем особую благодарность Ирине Куропаткиной и Анастасии Линник, оказавшим существенную помощь в сборе данных подкорпуса здоровых носителей языка.

2. Метод и материалы

2.1. Испытуемые

В настоящей работе были проанализированы рассказы 27 испытуемых, из которых 10 являлись здоровыми носителями русского языка (группа языковой нормы), а 17 — людьми с афазией разных типов (группа языковой патологии). Группа здоровых испытуемых состояла из 3 мужчин и 7 женщин, средний возраст — 63,7 лет. В группе людей с афазией было 7 мужчин и 10 женщин, средний возраст — 52,82 года.

2.2. Процедура

Длительность фильма о грушах (Chafe, 1980) составляет 5:55 минут. Фильм повествует о садовнике, собирающем груши, и о мальчике на велосипеде, который забирает одну из его корзин; развиваясь, история вовлекает в развертывание сюжета другие персонажи и события. В фильме нет устных диалогов, так что в его восприятии не задействованы процессы понимания речи (что делает его особенно ценным стимулом для изучения речевой патологии). Фильм был искусственно сконструирован для изучения различных аспектов нарративного дискурса; в частности, набор референтов, которые могут появиться в пересказах, — садовник, груши, мальчик и так далее — в большой степени предопределен заранее. Наличие общих референтов в дискурсивном пространстве предоставляет возможность сопоставления номинаций в различных нарративах.

Процедура записи нарратива заключалась в следующем. Сначала экспериментатор инструктировал испытуемого: «Мы исследуем, как люди рассказывают то, что видели. Посмотрите, пожалуйста, внимательно короткий фильм. И потом расскажите максимально подробно всё, что смогли запомнить». Испытуемый смотрел фильм, а затем экспериментатор приглашал в комнату слушателя (второго экспериментатора) и говорил: «Этот человек не видел фильма. Расскажите ему/ей о фильме максимально подробно. Так, чтобы он/она смог потом тоже рассказывать о фильме». На поведение слушателя накладывались определенные ограничения (именно поэтому на его месте не мог быть «наивный» испытуемый). Во время рассказа второй экспериментатор не должен был задавать развернутых наводящих вопросов (например, таких как «Куда они направились?», «Сколько было корзин?» и т. д.), но мог использовать общие реплики, провоцирующие дальнейший рассказ («Так», «Что дальше?» и т. п.). Во время эксперимента производилась аудио- и видеозапись.

Процедура была идентична в группах здоровых носителей и людей с афазией. Эксперимент с участием людей с афазией проходил на территории Центра патологии речи и нейрореабилитации (Москва) во время стационарного реабилитационного курса.

2.3. Анализ

Аудиозаписи полученных рассказов были затранскрибированы с помощью псевдофонетической записи. При анализе маркеров лексического затруднения мы следовали ряду ограничений. Так, в рамках данной работы не анализировались паузы, по определению преобладающие в не-беглой афазии (Murdoch, 2010) — ни заполненные, ни абсолютные. Следовательно, мы рассматривали только те случаи, когда лексические затруднения выражались вербально — попытками произнесения слова. Основными типами показателей затруднения порождения слова в нашем исследовании являлись дискурсивные маркеры (*это самое, как его*), в том числе предикативные (*скажем так, яблоки*), фальстарты (*гр = груши*) и самоисправления (*яблоки = нет, груши…*).

Случаи затруднений при порождении слова учитывались только тогда, когда они касались имен существительных.

Наконец, случаи затруднений при порождении слова учитывались только тогда, когда они касались имен существительных. Выбор предиката определяет всю структуру клаузы в целом, поэтому сбои при порождении и извлечении глагола могут повлечь за собой перестройку всего высказывания (Кибрик, Подлесская, 2009: 187–188). Соответственно, в основе таких сбоев могут лежать особые механизмы, рассмотрение которых не входило в цели настоящего исследования и не описывается используемой нами моделью.

Таким образом, в анализе учитывались вербально выраженные маркеры затруднения при порождении слов, обозначающих одного из присутствующих в дискурсивном пространстве историй о грушах референтов.К ним относятся: *садовник*, *сад*, *лестница*, *груши*, *корзина (корзины)*, *пастух*, *коза*, *мальчик*, *велосипед*, *девочка*, *шляпа*, *трое мальчиков*. Список включает только тех референтов, для которых в нашей выборке нарративов группы нормы и патологии встретилось как минимум два вербально выраженных маркера лексического затруднения.

3. Результаты и обсуждение

Мы рассматриваем наиболее характерные примеры вербально оформленных лексических затруднений. В представленных ниже примерах в скобках обозначена длительность пауз в секундах, фальстарты обозначаются дефисом (*мальчи- дети*). Элементы вербального оформления лексического затруднения выделены нижним подчеркиванием.

3.1. Дискурсивно-прагматический и лексико-семантический уровни

Поскольку ошибки семантики и выходного фонологического лексикона (то есть, неверного понимания значения слова и неправильного фонологического

представления) внешне могут не отличаться друг от друга и требуют специального тестирования (Howard, Gatehouse, 2006), эти два этапа порождения слова условно объединены в единый лексико-семантический уровень. Примеры дискурсивно-прагматических и лексико-семантических затруднений удобно рассматривать (ни в коей мере не объединяя эти принципиально разные явления) в одном разделе (3.1), так как иногда, в особенности для группы пациентов с афазией, однозначная интерпретация неочевидна.

3.1.1. Норма

(1) *другой* (1.5) *наверно фермер или кто-то там житель деревни, да*

В этом случае сбой номинации при последовательном разворачивании основной линии рассказа происходит на дискурсивно-прагматическом уровне. Говорящий эксплицитно (вербальные маркеры *наверно*, *кто-то там*) выражает свою неуверенность в адекватности выбранных им номинаций, при том что сами слова извлечены правильно. Неуверенность дополнительно маркируется вопросительным *да*, означающим запрос на подтверждение со стороны слушающего. Это проблема выбора, а не извлечения.

(2) *хозяин за- этот зс- э сборщик з- забрался высоко .* (2.2) *хозяин* (1.2) *вот тоже …*

Пример (2) представляет собой похожий случай: слова извлечены правильно, но говорящий не уверен в адекватности выбранной номинации, что выражается в двойном самоисправлении (*хозяин*, затем *сборщик*, затем снова *хозяин*).

(3) *что младший б- сын должен был приехать вот* (.) *за грушами*

Этот пример может быть проинтерпретирован двояко. Во-первых, аналогично (1) и (2)как неуверенность в адекватности номинации и ее сознательное исправление. Во-вторых, как ошибка на лексико-семантическом уровне: говорящий, начал извлекать устойчивое выражение *младший брат*, в контексте фильма маловероятно интерпретируемое как адекватная номинация, но оборвал высказывание (фальстарт), и далее последовало самоисправление (*сын*). См. также интерпретацию следующего примера:

(4) *но мальчи- дети все едят*

В этом случае трудно утверждать, с какого типа затруднением мы имеем дело.. Обе номинации (*мальчики* и *дети*) приемлемы в данном контексте, и неясно, что побудило говорящего оборвать слово (фальстарт *мальчи-*) и исправить его на другое (*дети*). Либо это непроизвольный ошибочный выбор слова из того же семантического поля, семантически связанного с целевым (тогда бы это оценивалось как сбой на лексико-семантическом уровне, связанный исключительно

с внутренними проблемами говорящего), либо сознательные колебания в выборе номинации на дискурсивно-прагматическом уровне.

3.1.2. Афазия

(5) *пасс-тан-чик* (0.3) *эта* (0.5) *э* (0.1) *эу* (0.1) *уносит* (0.8) *эээ ээ* (1.0) *ээ* (0.2) *овощи ой* (0.2) *мм фрукты*

Здесь замена семантически связанной с целевой, но некорректной номинации *овощи* на *фрукты* — это исправление ошибки на лексико-семантическом уровне, оформленное к тому же вербальным маркером осознания произошедшей ошибки — маркер *ой*.

(6) *яблоки это нет это как его*

В этом примере вербальными маркерами отмечены сами затруднения процесса извлечения.

(7) *собрал грибов* (0,2) *ой я уже наговорю незнамо чего* (0,7) *не знаю как*

Пример аналогичен предыдущему, но в этом случае осознанная ошибка (*грибов* вместо *груш*) порождает целые две полноценные предикативные конструкции, комментирующие затруднения говорящего на лексико-семантическом уровне.

(8) *отдал ребятам по по по целому яблоку яблоку по по по целой груше*

В этом примере самоисправление лексико-семантической ошибки дополнительно подчеркивается двумя случаями повторов предлога, эксплицирующими усилия по извлечению правильной лексемы .

(9) *бра-* (0.2) *ой* (0.2) *не братья а э друзья*

В примере (9) фальстарт с самоисправлением и вербальным маркером повторяет ошибку в той же номинации у нормы в примере (3). В этом случае вербальный маркер ошибки (*ой*) помогает интерпертировать пример именно как ошибку извлечения, а не затруднение на дискурсивно-прагматическом уровне, потому что в отличие от примера (3), обе номинации являются приемлемыми в контексте нарратива.

(10) *ээ* (0.3) *плоды* (0.1) *эта* (0.2) *груши*

Самоисправление, дополнено вербальным маркером (*эта*). В отличие от примера (6), в этом случае сложно понять, на каком именно уровне произошел сбой: как и в примере (9), обе номинации не являются неприемлемыми (в отличие от *овощей~фруктов* в (5)), однако в данном случае дискурсивный маркер не помогает уточнить уровень затруднения.

(11) *вот <u>й- яблоки</u> или <u>грибы ой не не гр- яблоки допустим да</u>*

В этом примере можно увидеть все три рассматриваемые нами вербальных способа оформления затруднения — дискурсивные маркеры, фальстарт с самоисправлением. При этом правильная номинация так и не будет извлечена; тем не менее, процесс лексической ошибки осознан говорящим (*грибы ой не не гр-*), равно как и паллиативный в данном случае выбор номинации *яблоки* (см. маркер *допустим*).

3.2. Фонологический уровень (выходной фонологический буфер)

3.2.1. Норма

В группе нормы, как и следовало ожидать, мы не обнаружили однозначных примеров затруднений на фонологическом уровне.

3.2.2. Афазия

(12) *ребята нашли его* (0.8) *э* (0.9) (1.1) *<u>сла- шляпу</u>*

Неправильно сформированная артикуляционная программа была осознана говорящим — произошел фальстарт, а затем самоисправление.

(13) *ве- веларис- вело-* (0,1) *са* (0,2) *пи* (0,5) *вело-ъ* (0,9) (шумный вдох) *съ- вело* (0,2) *си-* (0,1) *пед*

Ошибка на фонологическом уровне осознается испытуемым, поэтому в этом примере наблюдается серия фальстартов с самоисправлениями.

(14) *там <u>тв= твари= как его товарища</u> угостил*

В настоящем примере также представлены все рассматриваемые нами маркеры вербального затруднения: дискурсивный маркер и фальстарт с самоисправлением. Интересно отметить, что у того же самого говорящего маркер (*как его*), который он использует на протяжении всего рассказа, используется для оформления ошибок другого типа — см. пример (6).

4. Заключение

- Как и ожидалось, по сравнению с нормой, бóльшая часть собственно речевых сбоев, то есть семантических ошибок извлечения (лексико-семантический уровень), и в особенности ошибок оформления (фонологический уровень), происходит в нарративах группы афазии.

- Для группы нормы характерны преимущественно дискурсивно-прагматические ошибки, причем их прагматическая составляющая связана с особенностями экспериментального задания — пересказа фильма. Это мешает говорящему опираться на индивидуальный опыт, что является конституирующим признаком рассказа.
- Тем не менее, у обеих групп встречаются ошибки как на лексико-семантическом, так и на дискурсивно-прагматическом уровнях.
- Способы оформления дискурсивных и лексико-семантических ошибок (вербальные маркеры затруднений) пересекаются, поэтому в некоторых случаях трудно определить, на каком уровне произошел сбой.
- Рассмотренные нами способы оформления лексического затруднения встречаются и у одной, и у другой группы испытуемых при всех найденных нами типах ошибок.
- Способы оформления затруднений иногда могут помочь в интерпретации того, на каком уровне произошел сбой — см. примеры (9) vs (10). И наоборот, одни и те же маркеры могут сопровождать ошибки разных уровней — как в (6) и (14).

Таким образом, сопоставление результатов одного и того же дискурсивного задания для группы нормы и для группы людей, страдающих афазией, дает возможность увидеть отсутствие резких границ между речевыми сбоями, характерными для дискурса здоровых носителей языка, и ошибками, вызванными когнитивным дефицитом у людей с афазией — см. также Bergelson, Dragoy et al., 2010. Мы показали, что у нормы в конкретных точках (моментах) нарративного дискурса также имеют место локальные проявления когнитивного дефицита в рамках стандартных трудностей порождения правильных и уместных номинаций. Однако здоровые носители языка справляются с трудностями порождения нарратива, опираясь на различные языковые компетенции, например, рефлексируя относительно порождаемого ими текста, вступая в интеракцию со слушателем, эксплицируя свои рассуждения об уместности номинации. У людей с афазией эти компетенции так или иначе нарушены и потому представляют собой гораздо менее надежную опору. Так, возможность опоры на дополнительную информацию, способность эксплицировать свой логический вывод могут быть ограничены из-за обеднённости лексических и грамматических средств, доступных пациентам с афазией, и связанной с этим неуверенностью в своих возможностях.

В любом случае наши данные подтверждают валидность совмещения клинической и дискурсивной исследовательских перспектив. Для подкрепления этого общего вывода и в качестве будущей исследовательской программы будет необходимо провести не просто количественный анализ речевых сбоев в обеих группах, но и их разбиение по типам, а также анализ пауз как маркеров хезитации, анализ типов дискурсивных маркеров и просодии.

Литература

1. *Кибрик, А. А., Подлесская В. И.* (ред.). (2009) Рассказы о сновидениях: Корпусное исследование устного русского дискурса. М.: ЯСК.
2. *Коротаев Н. А., Кибрик А. А.* (2008). Иллокуция сообщения в устных рассказах: опыт корпусного исследования //Труды международной конференции «Корпусная лингвистика — 2008». СПб.: СПбГУ. — 214–220.
3. *Панов М. В.* (1968). Принципы социологического изучения русского языка. Русский язык и советское общество: В 4 кн. Кн. 1. М.
4. *Федорова О. В.* (2014). Экспериментальный анализ дискурса. М.: ЯСК.

References

1. *Ardila, A.* (2014), Aphasia Handbook, Florida International University, Miami.
2. *Benson, D. F., Ardila, A.* (1996), Aphasia: A clinical perspective, Oxford University Press, New York.
3. *Bergelson M. B., Dragoy O. V., Shklovsky V. M.* (2010), Telling a Story or Describing a Picture: Cognitive Differences and Similarities across Aphasic and Healthy Speakers, Proceedings of the Forth International Conference on Cognitive Science, Vol. 1, Tomsk State University, Tomsk, pp. 26–27.
4. *Chafe, W.* (ed.) (1980), The Pear Stories: Cognitive, cultural, and linguistic aspects of narrative production, Ablex, Norwood, NJ.
5. *Chapman S., Routledge C.* (eds.) (2009), Key Ideas in Linguistics and the Philosophy of Language, Edinburgh University Press, Edinburgh.
6. *Fedorova O. V.* (2014), Experimental study of discourse [Eksperimental'ny analiz diskursa], [YaSK], Moscow.
7. *Howard D., Gatehouse C.* (2006), Distinguishing semantic and lexical word retrieval deficits in people with aphasia. Aphasiology, 20(9), pp. 921–950.
8. *Kibrik A. A., Podlesskaya V. I.* (eds.) (2009), Night Dream Stories: A corpus study of spoken Russian discourse [Rasskazy o snovideniyakh: Korpusnoye issledovaniye ustnogo russkogo diskursa], [YaSK], Moscow.
9. *Korotaev N. A., Kibrik A. A.* (2008), Illocution of informing in oral narratives: Corpus study [Illokutsiya soobshcheniya v ustnykh rasskazakh: opyt korpusnogo issledovaniya], Proceedings of the international conference Corpus Linguistics-2008, [Trudy mezhdunarodnoy konferentsii "Korpusnaya lingvistika-2008], [SPbGU], pp. 214–220.
10. *Laine, M., & Martin, N.* (2006), Anomia: Theoretical and clinical aspects. Hove [England]: Psychology Press.
11. *Murdoch B. E.* (2010), Acquired Speech and Language Disorders, John Wiley & Sons, Chichester, Sussex, U. K.
12. *Nickels L.* (1997), Spoken word production and its breakdown in aphasia, Psychology Press, Hove, UK.
13. *Panov M. V.* (1968), Principles of the sociological analysis of Russian language [Printsipy sotsiologicheskogo izucheniya russkogo yazyka], In Russian Language and the Soviet Society [russkiy yazyk i sovetskoye obshchestvo], Moscow.

THE CASE OF RUSSIAN SUBJECT PRO IN MACHINE TRANSLATION SYSTEM

Bogdanov A. V. (abogdanov@abbyy.com)[1],
Gorbunova I. M. (igorbunova@abbyy.com)[1, 2]

[1]ABBYY, Moscow, Russia
[2]Russian State University for the Humanities, Moscow, Russia

This paper concerns a problem of Russian floating quantifiers (also known as semipredicatives) in machine translation. Floating quantifiers in Russian (such as *оба* 'both', *один* 'alone', *сам* 'on one's own' etc) are inclined for case, number and gender and agree in those categories with the subject of the minimal (finite) clause containing them. However, the case of a floating quantifier in an infinitive clause varies according to the type of PRO control applied and some other structural characteristics of the infinitive clause. This poses a problem for rule-based machine translation, to choose the correct case for the quantifier at synthesis, or to link it correctly to its antecedent at analysis. A model-based machine translation system, such as ABBYY Compreno, can handle the case choice problem, as this paper is to show.

Key words: subject, PRO, case marking, floating quantifiers, semipradicatives, machine translation, Russian

Introduction

This article deals with the problem of floating quantifiers in Russian from the perspective of ABBYY Compreno, a universal text analysis technology. Recently there has been a number of presentations concerning the information extraction features of the technology ([Anisimovich et al. 2012; Starostin et al. 2014; Bogdanov et al. 2014]). This article, however, deals more with the machine translation benefits that arise from the complete semantic-syntactic analysis of an input text, a task solved by Compreno.

1. Floating quantifiers and the case of Russian subject PRO

According to Babby, floating quantifiers are "adjectives that adjoin to VP and agree in case, gender and number with the subject of the minimal clause containing them". This can be illustrated by examples (1), (2), (3) and (4) (floating quantifier is marked nominative in (1), (3) and (4), but dative in (2); singular in (1–3), but plural in (4); masculine in (1–2), but feminine in (3), all in accordance to the features of the subject)

(1) *Я пришел сам*

(2) *Мне прийти самому?*

(3) *Она пришла сама*

(4) *Они пришли сами*

In Compreno syntactic parcer the notion of agreement is narrowed to a relation between two directly bound nodes in a tree and therefore in order to support the agreement floating quantifier is considered as moved from within subject NP, as in figure 1 (a syntactic tree in Compreno parcer for ex. (1))

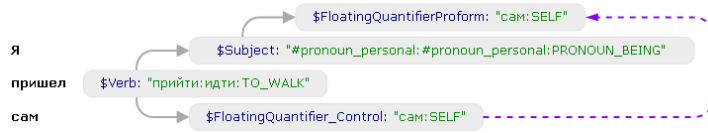


Fig. 1. Я пришел сам

The subject of infinitive clause in Russian is normally dative, as in (2). Indeed, in dependent infinitive clause with a PRO subject a dative floating quantifier can also be found, as in (5). However, the nominative case is sometimes the only option for a floating quantifier in infinitive clause, as in (6).

(5) *Он приказал нам прийти самим*

(6) *Он хочет прийти сам*

Amongst the numerous works on the subject of so-called second dative one can point out three main hypothesis about the nature of nominative and dative of floating quantifier in infinitive clause. Before turning to the description of the mechanism applied for the case choice in Compreno, we will give a brief account for those three hypothesis: universally local agreement ([Comrie 1974]), long-distance agreement for nominative and default dative assignment ([Franks 1990, 1995; Greenberg and Franks 1991]), as well as direct predication in subject control PRO constructions ([Babby 1998]).

Local agreement hypothesis appears in [Comrie 1974], the first paper to consider the problem of second dative. The idea is that the PRO of infinitive clause, whether lexically controlled or not, is assigned one of the cases—nominative or dative. Nominative is restricted to subject control PRO constructions and dative is a default case for the subject of infinitive clause. The syntactic structures for (5–6) are proposed as in (7–8):

(7) [*Он*] [*приказал нам*$_i$ [*PRO*$_{i,\,DAT}$ *прийти самим*$_{i,DAT}$]]

(8) [*Он*$_i$] [*хочет* [*PRO*$_{i,NOM}$ *прийти сам*$_{i,NOM}$]]

Long-distance agreement hypothesis was proposed by S. Franks. He claims that only a subject of a tensed CP infinitive clause can be assigned dative case, and PRO is essentially caseless. This claim is supported by the fact that most of the infinitives with an overt dative subject can take a tense auxiliary for future and past, as in (9) and its counterpart (10).

(9) *Куда нам поставить этот ящик?*

(10) *Куда нам было поставить этот ящик?*

Therefore, according to Franks, all the constructions where a subject of infinitive clause is overt and undoubtedly dative, are CPs, whereas dependent infinitives are IPs (as they are tenseless) and cannot assign dative to its subject. The nature of nominative and dative cases of floating quantifiers then are essentially different. Nominative case restricted to subject control PRO infinitive constructions is "transmitted" to the floating quantifier from the understood antecedent, with a long-distant agreement arising. Dative case on the other hand is a default case assigned to the sister to I' (the same rule as for the dative subject, but unlike the subject, a quantifier can be assigned the case directly without restriction to CPs).

Direct predication hypothesis was introduced in [Babby 1998]. Subject control PRO infinitive complement is viewed as a bare VP without a PRO, whereas other infinitives have a PRO which is assigned dative as a default case for the subject of infinitive. Floating quantifier thus receives the case form the nearest subject by agreement. The corresponding structures for (5–6) in this theory are (11–12)

(11) [*Он*] [*приказал нам*$_i$ [*PRO*$_{i,DAT}$ *прийти самим*$_{i,\ DAT}$]]

(12) [*Он*$_i$] [*хочет* [*прийти сам*$_{i.NOM}$]]

However, there is data that comes to conflict with each of the theories. First, it can be easily demonstrated that overt dative subject of infinitive clause is not bound to the tensed constructions. In (2) above one cannot add a tense auxiliary for future or past (unlike examples (9–10) borrowed from [Franks 1990]), and therefore one cannot argue that the construction in (2) is tensed. It contradicts the main argument of [Franks] that only a tensed CP can assign dative to the subject, and thus there is no consistent argument against the local agreement hypothesis[1].

Second, direct predication of [Babby 1998] can only account for infinitival sentential actants, but not for adjunct CPs such as that of (13), yet in (14) a nominative

[1] The same refers to the claim of [Fleisher 2006] that the overt dative NP in an infinitive clause is not subject to the infinitive clause but rather a subject to copula construction with infinitive complement.

case is acceptible (if not preferable) for the floating quantifier. As there can be no direct predication in (14), this hypothesis fails to explain this kind of nominative.

(13) *Я купил машину, чтобы ездить на работу самому*

(14) *Андрей слишком труслив, чтобы прийти сам*

This last pair of examples also pose a problem for the long-distance agreement hypothesis, as the infinitive clause in (14) is surely a CP (as it contains a conjunction in C), and therefore its subject must be assigned dative case, while nominative cannot be "transmitted" from above (CP must be blocking such a transmission).

Third, despite the claims of most of the cited authors, there are other cases accessible to the floating quantifiers in an infinitive clause, cf (15).

(15) *Меня просят прийти самого*

A problem therefore arises for the local agreement theory, because to explain examples like (15) one has to agree that the subject PRO of an infinitive must have a choice of three cases instead of two. When it comes to machine translation system, however, this latest problem appears to be the least of them all, as we will show in section 3.

2. PRO control in Compreno

In theory the control of infinitive PRO in Russian is dependent on theta-roles. Namely, the choice of the controller follows the hierarchy of [Jackendoff 1972: 43]

Patient > Addressee > Agent

This makes it difficult to build the control link without semantic analysis of the input text. For Compreno, however, this problem can be solved. As was already stated above, Compreno transfigures an input text into a semantic-syntactic tree, where each node is a notion given a package of grammatical information and diathesis description. Therefore, if some node is a parent to an infinitive clause, given all the information about the model of the lexical item in this node we can predict, what kind of control will be applied in the particular construction. For instance, consider (16–19).

(16) *Я пришел починить трубу*

(17) *Вы сказали мне починить трубу*

(18) *Меня прислали вам починить трубу*

(19) *Я был прислан вам починить трубу*

Compreno correctly coindexes the first person pronoun with the subject position of the infinitive clause in all the cases, cf fig. 2–5 respectively:

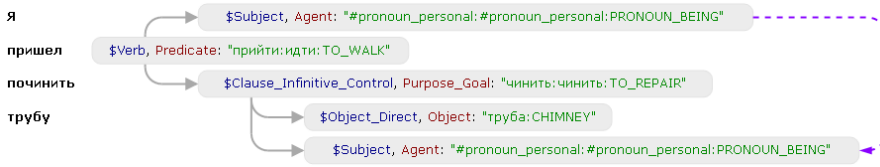


Fig. 2. Я пришел починить трубу

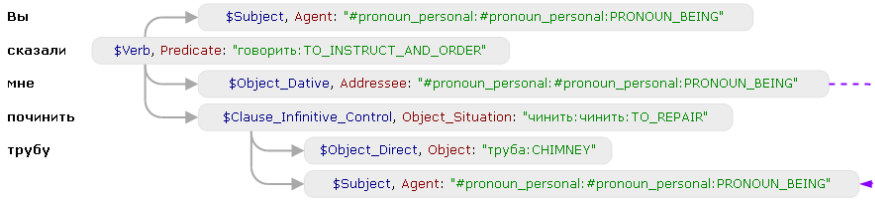


Fig. 3. Вы сказали мне починить трубу

Fig. 4. Меня прислали вам починить трубу

Fig. 5. Я был прислан вам починить трубу

As one can see, the non-tree links follow strictly the hierarchy rule mentioned above[2].

Technically this mechanism is comprised of two separate tools. First, each verbal lexical class is marked with a classifying flag that encodes the information about what type of control this verb can have when attaching infinitive (like SubjectControl, DirectObjectControl, DativeObjectControl etc.); second, when choosing the control type

[2] Object slot in Compreno system roughly corresponds to Patient role, referring to a general undergoer of the situation

not only those classifying flags are taken into consideration, but also the voice of the verbal node (like Active or Passive).

In the examples (18–19) above the verb прислать 'to send' is marked with a "DirectObjectControl" flag. For a verb with such classifying flag the control rule has but two options—the first is to build a link between its direct object and infinitive PRO if and only if the diathesis is active; the second is to build a link between its subject and infinitive PRO if and only if the diathesis is passive.

Having enumerated all possible combinations of classifying flags and information of the chosen diathesis it is not difficult to build a system of such non-tree rules as will be able to cope with all the PRO control infinitive constructions. It should be taken into consideration that there are verbs that can choose different adjuncts for antecedent in Active voice, cf.:

(20) *просить мальчика сделать что-то*

(21) *просить у мальчика сделать что-то*

Although the number of combinations is therefore quite large, this system in general is nevertheless pretty straightforward.

3. Floating quantifiers and the case choice in Compreno

Floating quantifiers, as already shown in section 1 (fig.1, 6), are moved from within NP in Compreno syntactic structure and agree in case, number and gender with the parent node before movement. So for the floating quantifier to be marked with case X the subject of the minimal clause containing it has to be assigned the same case X. This is consistent with the local agreement hypothesis rather than any other.

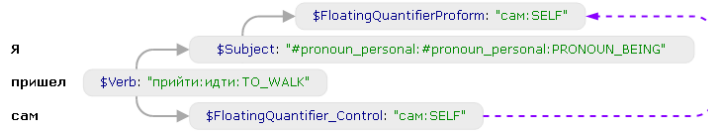


Fig. 6. Я пришел сам

This implies that a subject of an infinitive clause can be nominative (as in (22)), dative (as in (23)) or accusative (as in (24)).

(22) *Я хочу починить трубу сам*

(23) *Мне сказали починить трубу самому*

(24) *Меня просят починить трубу самого*

Moreover, there is a number of cases where the floating quantifier is unacceptable or at least dubious. Such are the instances where the PRO of the infinitive clause is coindexed with instrumental NP, as in (25).

(25) *Правительством планируется восстановить разрушенные территории *само/*самим*

In Compreno this problem is solved as follows. Every infinitive node of the tree (after the tree is built and all the non-tree links are established) is assigned a special flag that encodes information for the type of control applied in the particular structure. Let us call it TypeOfPRO flag. Restricting the subject-predicate relation, which is represented as one arc in the tree, we assign a certain case to the subject according to the flag of the parent node. Consider semantic-syntactic trees for (22–23), figures 7–8 respectively.

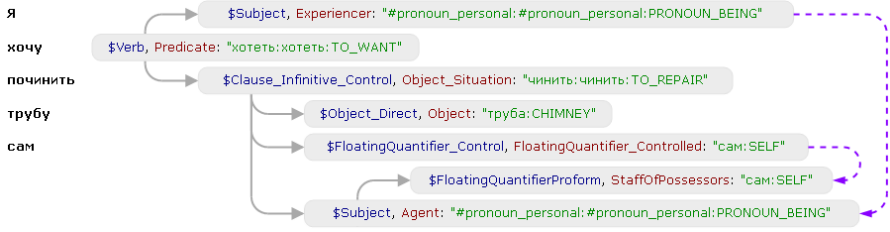


Fig. 7. Я хочу починить трубу сам

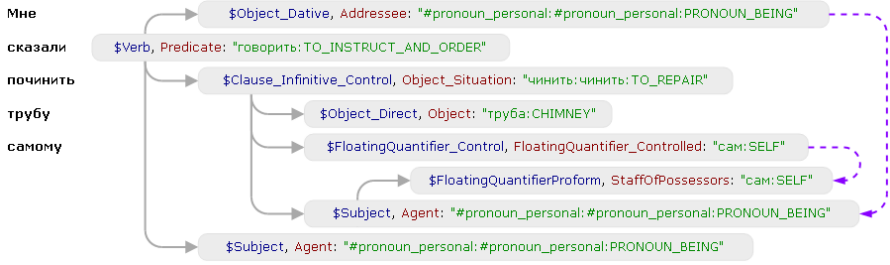


Fig. 8. Мне сказали починить трубу самому

In (22), figure 7, the PRO of infinitive clause is controlled by the subject of the matrix predicate. Due to it, the infinitive node bares the NominativePRO flag and its PRO is assigned nominative case. It transmits nominative to the floating quantifier before movement, so that the moved quantifier is also marked nominative. This makes example (26) with dative case invalid.

(26) **Я хочу починить трубу самому*

In (23), figure 8, the PRO of infinitive clause is controlled by the dative object of the matrix predicate, so the infinitive node bares DativePRO flag. Its PRO is assigned dative case and transmits dative to the floating quantifier by agreement. Thus (27) with nominative is also analyzed as invalid

(27) **Мне сказали починить трубу сам*

By the same principle object control PRO as in (24) can be assigned accusative and allow for the floating quantifier to be accusative via agreement. As for (25), the infinitive node is marked with MarginalPRO flag, which means that its PRO is controlled in such a way that some syntactic transformations are blocked inside this clause. Floating quantifier movement is one of those blocked transformations, hence unacceptability of (25).

4. Further applications of the TypeOfPRO flag

The mechanism illustrated above has several other applications apart from the case choice for the floating quantifiers. It has been noticed before, that subject control PRO infinitive constructions can take a short form adjective as a complement (28), whereas object control PRO infinitive constructions cannot (29)

(28) *Я должен/хочу быть красив/красивым*

(29) *Мне хочется быть *красив/красивым*

As there are lexical items without full form it is crucial for the machine translation to choose a synonymous lexical item for constructions with object control PRO infinitive, cf (30–31).

(30) *Я должен быть рад*

(31) **Мне хочется быть рад*

In Compreno it is simply done by applying the TypeOfPRO flag on the infinitive node to restrict form choice in the complement node. It is according to this test that the PRO in constructions such as (25) are assigned dative in Compreno: short form of adjective is unacceptable as a complement in such constructions, cf. (32)

(32) *Правительством планируется быть *компетентно/компетентным*

5. Conclusion

Although a model-based approach to machine translation is known to be relatively labour-intensive, it looks more promising when interpreting and translating such complex structures as those with floating quantifiers in infinitive clauses. For the analysis of those constructions it seems more reasonable to follow the local agreement hypothesis and assign case to PROs, however intuitively dubious that may be.

References

1. *Anisimovich K. V., Druzhkin K. Ju., Minlos F. R., Petrova M. A., Selegey V. P., Zuev K. A.* (2012), Syntactic and semantic parser based on ABBYY Compreno linguistic technologies, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog" [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii Dialog"], Bekasovo, pp. 90–103.
2. *Babby, L. H.* (1998), Subject control as direct predication: Evidence fromRussian, Annual Workshop on Formal Approaches to Slavic Linguistics: The Connecticut Meeting, pp. 17–37.
3. *Bogdanov A. V., Dzhumaev S. S., Skorinkin D. A., Starostin A. S.* (2014), Anaphora analysis based on ABBYY Compreno linguistic technologies, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog" [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog"], Bekasovo, pp. 89–102.
4. *Comrie B.* (1974), The second dative: A transformational approach, Slavic Transformational Syntax, No 10, pp. 123–150
5. *Fleisher N.* (2006), Russian dative subjects, case, and control, Ms., University of California, Berkley
6. *Franks S.* (1990), Case, Configuration and Argumenthood: Reflections on the Second Dative', Russian Linguistics, Vol. 14, pp.231–254.
7. *Franks S.* (1995), Parameters of Slavic Morphosyntax, Oxford University Press, Oxford.
8. *Greenberg G. A., Franks S.* (1991), A parametric approach to dative subjects and the second dative in Slavic, Slavic and East European Journal, Vol 35, pp. 71–97.
9. *Jackendoff R* (1972), Semantic Interpretation in Generative Grammar, MIT Press, Cambridge, MA.
10. *Starostin A. S., Smurov I. M., Stepanova M. E.* (2014), A production system for information extraction based on complete syntactic-semantic analysis, available at: http://www.dialog-21.ru/digests/dialog2014/materials/pdf/StarostinAS.full.pdf.

СЕМАНТИЧЕСКИЙ АНАЛИЗ И ОТВЕТЫ НА ВОПРОСЫ: СИСТЕМА В СТАДИИ РАЗРАБОТКИ

Богуславский И. М., Диконов В. Г., Иомдин Л. Л., Лазурский А. В., Сизов В. Г., Тимошенко С. П.

Институт проблем передачи информации РАН им. А. А. Харкевича, Москва, Россия

В статье представлена система семантического анализа и вопросно-ответная система, реализованная на ее основе. Предметной областью являются новости про футбол. На входе система получает вопрос на естественном языке, в качестве ответа выдаёт элемент из базы данных. Модуль семантического анализа лингвистического процессора ЭТАП-3 строит для каждого предложения семантическую структуру, представляющую собой набор троек вида **`семантическое_отношение(индивид,индивид)`**. Семантические отношения и индивиды, из которых состоит семантическая структура, соответствуют элементам онтологии, которая таким образом становится функциональным аналогом словаря для «семантического языка». Семантические структуры предложений одного текста объединяются благодаря установлению кореференции между объектами и конвертируются в OWL-документ, использующийся далее в качестве базы данных. В базу данных также помещаются фоновые сведения из базы индивидов о конкретных командах, футболистах, матчах. Благодаря этому становится возможным находить ответ на вопрос, используя информацию, содержащуюся не только в разных предложениях текста, но и в базе индивидов. Так, если пользователь задал вопрос *Какая команда нанесла поражение чемпиону Испании?*, а мы располагаем текстом, в котором сообщается, что *Подопечные Слуцкого обыграли мадридский «Атлетико»*, то система установит соответствие между вопросом и этим текстом и даст правильный ответ: *ЦСКА*. Семантическая структура, полученная из вопроса, конвертируется в SPARQL-запрос к базе данных. На данный момент все части системы функционируют, работа находится в стадии отладки.

Ключевые слова: глубокий семантический анализ, семантический словарь, онтология, вопросно-ответная система, кореферентность

SEMANTIC ANALYSIS AND QUESTION ANSWERING: A SYSTEM UNDER DEVELOPMENT[1]

Igor Boguslavsky, Vyacheslav Dikonov, Leonid Iomdin, Alexander Lazursky, Victor Sizov, Svetlana Timoshenko

A. A. Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia

The paper presents a system of semantic analysis and a question answering system implemented on its basis for a specific subject domain: (European) football match news. As input, the system obtains a natural language question (in Russian), which it answers with an element (or elements) from the repository of individuals. The core part of the system is the semantic analyzer of natural language texts. For each sentence of the text processed, the special semantic analysis component of ETAP-3 linguistic processor constructs a semantic structure, which consists of a set of triples of the type **`semantic_relation (individual,individual)`**. Semantic relations and individuals constituting this structure correspond to the elements of the ontology, which can thus be viewed as a functional analogue of a dictionary for the semantic language. Semantic structures of sentences belonging to a particular text are integrated thanks to coreference and anaphora resolution and converted into an OWL-document, which is later used as a database. This database is supplemented by background knowledge from the repository of individuals concerning specific teams, football players, and games. Thanks to this resource, we are able to find an answer to the question using not only the data contained in different sentences of the text but also in the repository of individuals. If the user asks "*What team defeated the champion of Spain?"* while we have a text reporting that "*Slutsky's players outplayed Atletico Madrid*" then the system will establish the correspondence with the question, the text, and the records in the depository of individuals, and will come with the correct answer "CSKA". The semantic structure obtained from the natural language question is converted into a SPARQL query addressed to the database. Currently, all parts of the system are operating in the test mode.

Key words: deep semantic analysis, semantic dictionary, ontology, question answering, coreference

1. Introductory remarks

Semantics is the most essential and probably the most complex component of the full model of natural language. So far, the task of semantic analysis is far from being

[1] The paper is partially funded by the Russian Humanitarian Scientific Foundation, grants No. 13-04-0043 and 15-04-00562, and the Russian Foundation of Basic Research, grant No. 15-06-09208. The authors express their gratitude to both Foundations.

fulfilled, even though numerous attempts to achieve this goal using diverse approaches have been made. Many researchers view the task of semantic analysis in tagging the text by semantic elements, such as WordNet synsets, ontology classes or individuals, semantic roles, or FrameNet frames (Shi and Mihalcea, 2004; Coppola, Moschitti, 2010; Azmeh et al., 2011). In a different approach, many authors attempt to translate natural language sentences into a logical language (see e.g. Bos, 2008, 2011; Copestake et al., 2006, Allen et al., 2008). On the other hand, many papers focused on semantic analysis tend to use, in addition to linguistic data, also background information contained in the ontology. This approach, primarily pursued in the OntoSem project (Nirenburg, Raskin, 2004; Akshay Java et al., 2007; Raskin, Taylor, 2010; Raskin et al., 2010; Nirenburg, McShane, 2012) is helpful in tasks of lexical and/or syntactic disambiguation and can be used in deducing all sorts of inferences which contribute to deeper and more comprehensive understanding. A series of articles written by the Spanish FuncGram group who works in the framework of Lexical Constructional Model develop a semantic database, used for inferences (see e.g. Mairal Usón, Periñán-Pascual, 2009; Periñán-Pascual, Arcas-Túnez, 2010a, b; Periñán-Pascual, Mairal Usón, 2010). Automatic semantic analyzers are actively developed within the machine learning paradigm, especially under supervised learning (cf. Ge and Mooney, 2005; Poon and Domingos, 2009; Clarke et al., 2010; Titov and Klementiev, 2011; Liang et al., 2011). An obvious obstacle here is the lack of sufficiently large semantically tagged corpora. It should be added that some semantic parsers combine mixed technique: machine learning and linguistic rules (Moldovan et al., 2010).

Unlike these last ones, our analyzer of Russian texts is strictly rule-based, which seemingly contradicts the current trend. Our choice of strategy is based on two considerations. First, there exist no corpora annotated with the kind of structure we are interested in. Once we construct our analyzer, it will open the possibility to develop such a corpus, which could then be used for refining and evaluating the analyzer, as well as for developing other semantic parsers. The second, and more important, reason for our no statistics approach is our firm belief that the modelling of real understanding of texts requires knowledge-intensive methods (for details, see Boguslavsky 2011).

Most researchers agree that the goal of constructing a broad-coverage system of deep semantic analysis is currently unachievable. There are two possible ways to produce such a system: (1) start with an extensive shallow understanding system and gradually deepen it, as suggested in (Riloff, 1999), or else begin with a small-scale deep understanding system and gradually broaden the coverage (Mueller, 2006). We advocate the latter approach. Our semantic analyzer prototype is aimed at deep understanding of texts belonging to a restricted subject domain: football match news. To achieve this, we need to use not just linguistic knowledge but subject domain knowledge which is provided by the ontology.

The football subject domain has repeatedly attracted the attention of research groups working with ontologies. Several ontologies have been built, which contain a rather comprehensive nomenclature of football terms. These include the SWAN Soccer Ontology by DERI (http://sw.deri.org/ knud/swan/ontologies/soccer), Ranwez's ontology (http://www.lgi2p.ema.fr/~ranwezs/ontologies/soccerV2.0.daml), the sports fragment of the OpenCyc Ontology (http://sw.opencyc.org/2009/04/07/concept/en/Soccer), the sports fragments in the DAML repository (http://www.cse.sc.edu/~dukle/

ontologies/football-ont.daml) (Dukle, 2003), the soccer ontology of the i3media project (Bouayad-Agha et al. 2011), and the Kiktionary Ontology (kiktionary.de). Yet, neither of these ontologies meets the needs of our project so that we had to develop our own ontology. The reason is that the basic application of the ontologies listed above is annotation of texts or multimedia objects aimed at facilitating database search. Typical examples are Ranwez ontology intended for video file annotation, and the paper (Tsinaraki et al., 2005), who discuss methods of using the ontology for semantic indexation of audiovisual material. The Kiktionary ontology is a trilingual dictionary of football terms arranged into several hierarchies, frames, and scenes. Somewhat different is the i3media ontology intended for the generation of summaries of football matches from statistical reports of these matches. The goals determine the contents of these ontologies. Information on each concept supplied therein basically confines to the concept's referral to a higher class of hierarchy and the list of equivalents in one or more languages. To give an example, the Goal concept is presented in Ranwez in the following way (using OWL):

```
<rdfs:Class rdf:ID="Goal">
<rdfs:subClassOf rdf:resource="#Stoppage"/>
<rdfs:label xml:lang="fr"> But
</rdfs:label>
<rdfs:label xml:lang="en"> Goal
</rdfs:label>
</rdfs:Class>
```

This description reads as follows: «Concept Goal is an element of the Stoppage class. In English, the corresponding term is *goal*, and in French it is *but*». Other ontologies provide similar information. This level of explication is quite sufficient for tagging football news reports with ontological concepts. However, it offers no understanding on what this event is about and does not allow drawing any inferences that require such understanding. A more detailed description of this concept given in our OntoEtap ontology is shown below, in section 3.

An important component of the ontology is the set of class instances, which may be many times greater than the set of classes and properties. For example, the i3media project ontology contains 47 classes, 50 properties, and ca. 70 thousand instances. Often, ontologies are automatically populated with data scraped from web pages (Bouayad-Agha et al., 2011; Dukle, 2005) by means of specially designed programs.

In our project, the construction of a complete ontology of football and its population with instances is not our main focus of attention. These tasks were solved previously, and technologies of populating the ontology by different methods, including automatic ones, are known. Our main route of advancement is to learn how to extract implicit knowledge from texts and make inferences based on this knowledge. The extent of ontological coverage plays no essential role in the development of methods that allow for such reasoning. Much more important is to understand exactly which knowledge the system needs to solve this task, to learn how to produce this knowledge and operate it. This approach entails the requirements that should be met by our resources to be, and the order of action. In accordance with this approach, we decided to start

by constructing a fragment of an ontology which should not be big but which should contain knowledge that allows the system to fetch the implicit information extractable from texts. Should the experiments performed in this restricted area prove successful, we could proceed with the expansion of the area and move towards a fuller coverage of the subject domain. This is why we venture to report on this project without waiting for its maturity and high recall that would allow for its full-scale evaluation.

More specifically, our semantic analyzer strives to advance in the following directions:

(i) The use of extralinguistic knowledge in addition to linguistic data presented in the dictionary and the grammar. Extralinguistic knowledge is stored in two repositories: the Ontology and the Repository of Individuals. While the Ontology stores hierarchically arranged information on concepts and their properties, the Repository of Individuals accumulates data on individual objects (like Moscow) or situations (like 2014 FIFA World Cup).

(ii) Explicit presentation of word meanings for inference purposes. We proceed from the assumption that the depth of understanding is growing with the number of inferences we can draw from the text. In many cases, a detailed description of word meanings helps produce additional conclusions and in this achieve a deeper understanding.

(iii) Going beyond the sentence boundary. Normally, syntactic and semantic analysis of text is limited to one sentence, so that it is impossible to look from the sentence processed to a neighbouring one. It is however a serious obstacle for many tasks. Importantly, going beyond the sentence is essential for finding antecedents of pronouns which are very often located in one of the preceding sentences. We will also show below that in order to answer relevant questions to the text it may be essential to resort to the material of several sentences.

Our analyzer is built on the basis of ETAP-3 linguistic processor as its new module. For more details see (Iomdin et al., 2012), (Boguslavsky et al., 2013).

The paper is written according to the following plan. Section 2 will be focused on OntoEtap ontology, covering the three issues just listed. Section 3 will show how word/phrase meanings are represented. Section 4 will discuss how the context is treated. Section 5 will give a few detailed examples demonstrating the system implementation in ETAP-3. Although full-fledged system evaluation in terms of precision and coverage is not possible at this stage, it is nevertheless desirable to make sure that the information produced by the system can work. Therefore, we decided to carry out some restricted experiments in the question-answering scenario in which this information would come in handy. These experiments are described in Section 6. In section 7 we will outline directions of future work.

2. OntoEtap Ontology

The ontology is a natural intermediary link that connects natural language with extralingusitc knowledge and the processes with which this knowledge is manipulated. There are two distinctly different sorts of such knowledge.

One sort is Ontology proper. It consists of a hierarchically arranged set of concepts with formal properties assigned to them. For instance, the class concept **City** denotes the class of cities, which is a subclass of the class **GeopoliticalArea**; in its turn, **City** has a subclass **CapitalCity**. The concept **City** is assigned such properties as the country where the city is situated, its population, area, geographic coordinates etc. All properties of a class are inherited by its subclasses. Many concepts refer to more than one class, thus inheriting properties coming from different sources.

The other sort of extralinguistic knowledge is the Repository of individuals, which contains information on individual objects or situations that are concrete instances of Ontology concepts. Individual objects are e.g. Moscow, France, the Thames, or Cervantes, and individual situations include World War II, Sochi 2014 Winter Olympics, or Yesterday's match between Arsenal and Manchester United.

We have chosen OWL (Web Ontology Language) to present both the Ontology and the Repository of Individuals, because this language is common for ontology developers and ontological semantics community and because a number of useful tools are available for OWL manipulation (such as editors, reasoners, or workplaces).

Our OntoEtap ontology has two sources. One source is the well-known upper/middle ontology SUMO, or Suggested Upper Merged Ontology (www.ontologyportal.org), rewritten in OWL. SUMO is an open formal ontology formulated in the first order predicate logic language. It contains references to WordNet and involves several restricted subject domain ontologies, covering jointly about 20,000 concepts and 60,000 axioms. The other source is our own small ontology of the football news domain, written in OWL using SWRL rules. The resulting OntoEtap ontology is maintained in Protegé enviroment. Today (April 2015), OntoEtap contains 10,963 classes and 5,587 individuals.

3. Representation of Word Meanings

The role of Ontology in our project is twofold. On the one hand, it is the source of structured world knowledge. On the other hand, it serves as a metalanguage of semantic representations. Concepts, individuals and their formal properties are viewed as semantic elements with which semantic structures are built. All (non-functional) Russian words are interpreted in terms of these semantic elements.

To give an example, all the diverse designations of a victory/defeat in a football match, such as *победить* 'defeat', *выиграть* 'win', *переиграть* 'outplay, beat', *разгромить* 'rout', *заработать 3 очка* 'win 3 points'; *проиграть* 'lose', *уступить* 'concede', *потерпеть поражение* 'suffer a defeat', *оказаться сильнее/слабее* 'be stronger/weaker against' etc. are eventually represented by one general concept **WinEvent**. The difference between winning and losing is not reflected at the level of concepts but manifested in the way in which the **WinEvent** concept's roles **hasWinner** and **hasLoser** are implemented.

In addition to words that correspond to a single concept (*победить* 'win' to **WinEvent**, *футболист* 'footballer' to **FootballPlayer** etc.) there are words whose meaning is represented by a structure formed with several concepts. The word

генерал '(army) general' may refer to a military rank or to a human who has this rank. The first sense of the word is tantamount to the concept **GeneralRank**, which belongs to the class **MilitaryRank**, while the second sense corresponds to the structure **hasRole(Human,GeneralRank)**. A different example: the Ontology includes concepts corresponding to animal species, like **Lion**, but it would be redundant to introduce special concepts for words like *львица* 'lioness' or *львенок* 'lion cub'. Such words are best represented with structures that explain their senses with the concept **Lion**: **hasGender(Lion,female)** and **developmentalForm(Lion,NonFullyFormed)**.

Rather often, the link between the natural language and the ontology is understood as mapping NL expressions to ontology nodes. Yet, natural language expressions do not always have a fixed ontological correlate. For instance, expressions like *local team* correspond to different ontological individuals depending on the context. For such cases, special rules of contextual interpretation should be written.

Complex events are described by means of multi-propositional explications (scripts). For example, the central event that happens in a football match—scoring a goal—is supplied with a description that interprets this event as a script formed with three sub-events that are causally related with each other[2]: a **GoalEvent** situation takes place if:

(1) Player who belongs to Team-1 hits the ball in the direction of the GoalArea of Team-2.
(2) As a result, the ball is located in the GoalArea of Team-2.
(3) As a result, Team-1's score increases by one.

Due to the decomposition of the goal concept into several components we can generate an adequate representation of expressions in which these components come into play. For instance, phrases like *забить гол головой (с 20 метров)* 'score a goal with one's head (from 20 meters)' may only be understood if we take into account that a goal event includes hitting the ball (cf. *hitting the ball with the head (from 20 meters)*).

Besides, the description of an event with the help of a script opens a possibility of treating certain kinds of metonymy. It allows one to identify the event even in cases when the text only mentions a part of the relevant components or even a single crucial component. The following sentences do not contain even the word *гол* 'goal' but obviously refer to this sort of event: (1) *Уже на второй минуте вратарь достал мяч из ворот* 'As early as on the second minute, the goalkeeper took the ball out of the goal'; (2) *Месси отправил мяч в верхний левый угол ворот*. ‹Messi sent the ball to the upper left corner of the goal›.

Let us see how these sentences are processed. First, we produce the Basic Semantic Structure (BSemS), which conveys the basic meaning of the sentence in terms of ontological units. BSemSs of sentences (1) and (2) are shown in Fig. 1 and Fig. 2, respectively.

---

[2] To save space, we are not reproducing the formal definition. Many examples of using this language for word definitions can be found in Boguslavsky et al. 2013.

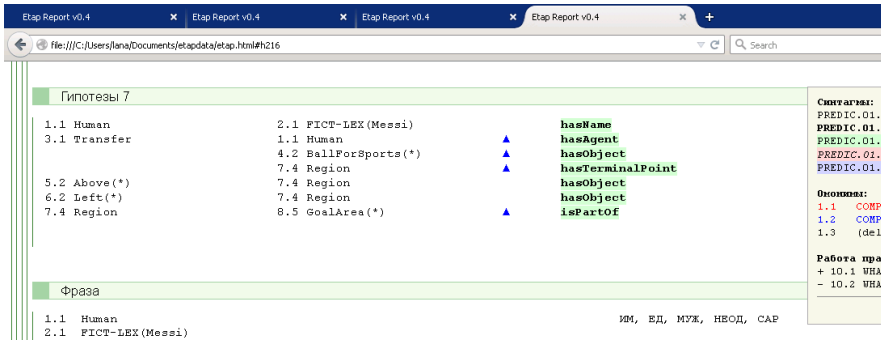


Fig. 1. Basic Semantic Structure (BSemS) for sentence (1)

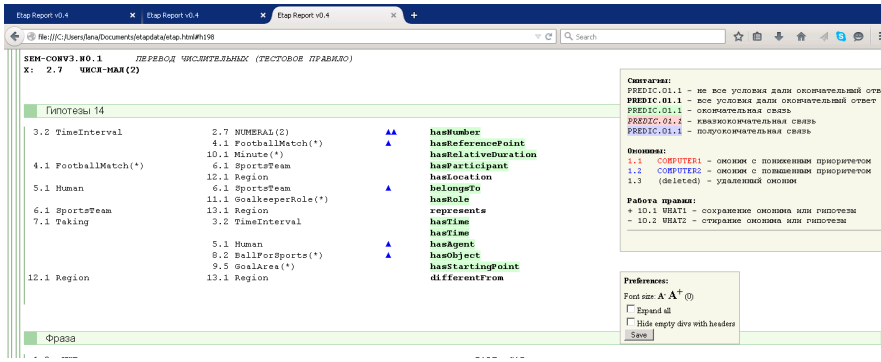


Fig. 2. Basic Semantic Structure (BSemS) for sentence (2)

After that, inference rules are applied, if need be. These rules make various kinds of inferences and have different degree of generality. Among them, there are (a) concept-centered rules, such as effect and precondition rules, which are specific to particular concepts, and (b) more general common sense rules. Both kinds of rules are needed for processing sentences (1) and (2). They are written in one of two formalisms—SWRL (Semantic Web Rule Language) or FORET (ETAP rules language). Below, we formulate these rules in plain words to facilitate understanding by the readers.

(i) if a physical object is taken out of some place, it was previously located in this place (the precondition of the TakeOut event);

(ii) if a ball is located in the goal area, this is the result of the goal (the effect of the GoalEvent; this rule is a kind of abductive inference).

(iii) if a physical object is moved to a place, it becomes located in this place (the effect of the Transfer event)

Rules (i)–(ii) are applied to sentence (1) and generate four new triples boxed in Fig. 1. Sentence (2) requires rules (iii) and (ii), but before (ii) could be applied, a common sense rule (iv) should have been applied:

(iv) if an object is located in place A which is part of place B, it is located in place B (transitivity of the hasLocation relation).

After these rules are applied to sentence (2), its BSemS is supplemented with new triples (boldfaced in Fig. 2), which show that the GoalEvent has taken place.

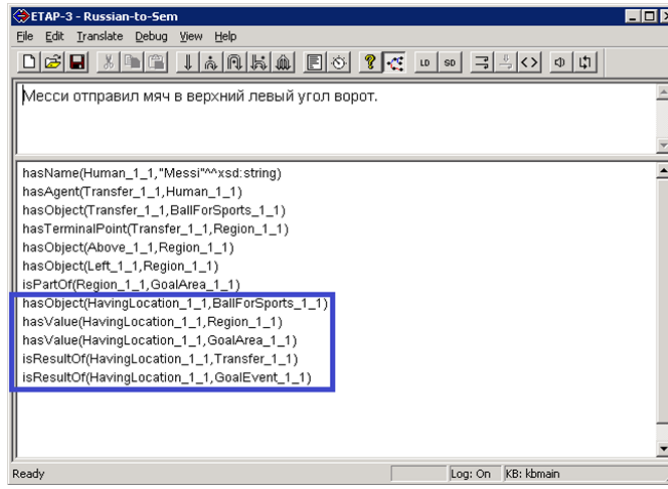


Fig. 3. Inferred semantic structure of sentence (1)

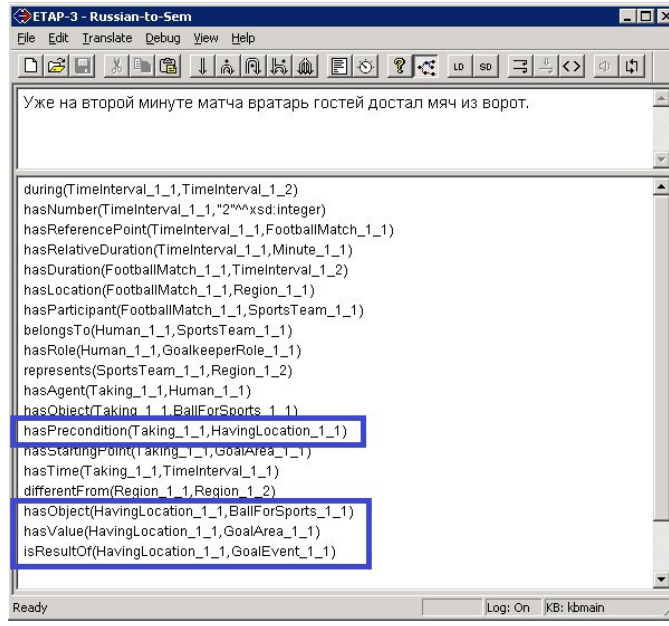


Fig. 4. Inferred semantic structure of sentence (2)

Thus, our analyzer not only interprets complex concepts in terms of simpler ones, but also makes certain kinds of inference.

4. From Sentence to Text

We have mentioned above that the semantic analyzer must operate across sentence boundaries. For a question answering system based on such an analyzer (see Section 6 below for details), this requirement can be reformulated as follows: ideally, the system should be able to find an answer to any question even if it involves collecting the material scattered in different sentences.

The material for the experimental question answering system has been a collection of football news published on the lenta.ru portal. Every message consists of 5 to 10 sentences and normally represents one event. When semantic structures of different sentences are merged into a single description, one of the most difficult problems to be solved is establishing the coreference (or absence thereof) between the objects mentioned in one (or, especially, more) sentences. In our texts, this problem often arises with different names of matches, teams, and players, because the authors tend to avoid repeating the same designations in adjacent sentences. A typical example could be seen in the message *ЦСКА и «Торпедо» встретились в Испании в матче турнира Pinatar Cup. Победу в столичном дерби одержали черно-белые* 'CSKA and Torpedo met in Spain in the Pinatar Cup tournament. In the capital derby, the black-and-whites won'.

In many cases, different designations could be identified as referring to the same entities with the help of the Repository of Individuals. For instance, the entry for the Torpedo team contains the information **`hasNickname "Cherno-belye"`** ('black and whites'). Yet we do not have a general solution of the problem, so we need to look for partial solutions. Considering the fact that news messages normally (though not always) cover only one sports event, we assume that different mentions of an event within a message are coreferent if no evidence to the contrary are available. Pointing to different times, locations, or lists of participants may form such evidence.

Suppose for example that we need to find an answer to the following question: *Как сыграли «Бавария» и «Реал Мадрид»?* 'How did Bavaria and Real Madrid play?'

In our news collection, we have the following message fragment:

(3) *Мюнхенская «Бавария» обыграла мадридский «Реал» в первом полуфинальном матче футбольной Лиги чемпионов. Встреча, проходившая на стадионе «Альянц-Арена» в Мюнхене, завершилась со счетом 2:1.*
'Munich Bavaria beat Real Madrid in the first semifinal match of the football Champions League. The meeting, held at the stadium "Allianz Arena" in Munich, ended with the score 2:1'.

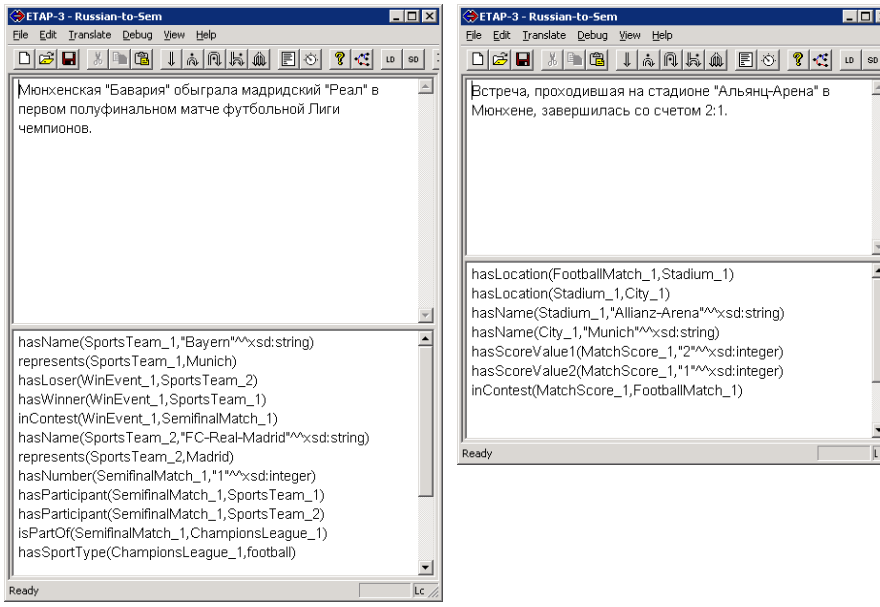
In order to extract the full answer to this question from the given text (namely, that *«Бавария» победила «Реал» со счетом 2:1* 'Bavaria beat Real with the score 2:1'), we need to understand that both sentences report on the same match. This could be done because there is no evidence that the matches are different.

5. Case studies

Below, we will demonstrate two techniques of semantic analyzer operation.

5.1. Building a semantic structure of the message with semantic structures of individual sentences

Let us return to text fragment (3) considered in Section 4 above. The semantic structures of the two sentences are presented in Fig. 5a-b. Both figures are screenshots of ETAP-3 linguistic processor operation.



a b

Fig. 5. Semantic Structures of Text Fragment (3): a—first sentence, b—second sentence

The structure in Fig. 5a can be "read" in the following way: "The team SportsTeam 1 is called Bayern (which is German for Bavaria) and represents Munich (lines 1–2). It lost to team SportsTeam 2 in a semifinal match SemifinalMatch_1 (lines 3–5). Team SportsTeam 2 is called FC-Real-Madrid and represents Madrid (line 6–7). The semifinal match in which these teams participated is part of the Champions League and has sequence number 1 (lines 8–12).

The structure in Fig. 5b says that some football match FootballMatch_1 took place in some stadium Stadium_1 located in some city City_1 (lines 1–2), this stadium is called Allianz-Arena and the city is called Munich (lines 3–4), the score for one team ScoreValue1 in MatchScore_1 was 2 (line 5), the score for the other team

ScoreValue2 in the same MatchScore_1 was 1 (line 6), and this MatchScore_1 was reached in football match FootballMatch_1 (line 7).

The two semantic structures are merged into one integral structure of the text, in which the coreference is established between the event SemifinalMatch_1 from the first sentence and the event FootballMatch_1 from the second sentence. Now the structure is ready to be used in answering the question.

5.2. Addressing the information on individuals stored in the Repository of Individuals

As was already mentioned, the semantic analyzer can access the information on individuals that is absent from the text and is only represented in world knowledge. This access allows us to state the referential identity of expressions which look extremely different. In its turn, the question answering system enhances its potential and becomes able to answer a much broader range of questions. If, for example, the user asks (4) *Какая команда нанесла поражение чемпиону Испании?* 'What team defeated the champion of Spain?' while we have a text reporting that (5) *Подопечные Слуцкого обыграли мадридский «Атлетико»* 'Slutsky's players outplayed Atletico Madrid' then the system will establish the correspondence between the question, the text, and the records in the depository of individuals, and will come with the correct answer "CSKA". This inference relies upon the background knowledge that Slutsky is the trainer of CSKA and that Atletico Madrid won La Liga 2013–14 tounrament, which is the Championship of Spain. All these data can be seen in Fig. 6–7 below that represent the ETAP-3 semantic analysis operation screenshots, and in Fig. 8a–b that represents information from the depository of individuals.

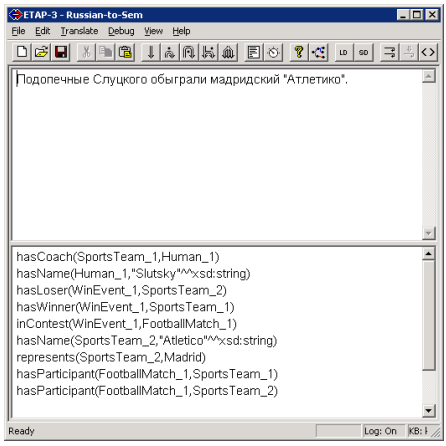


Fig. 6. Semantic structure of the question

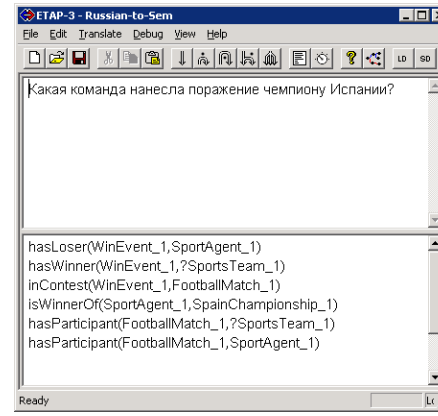
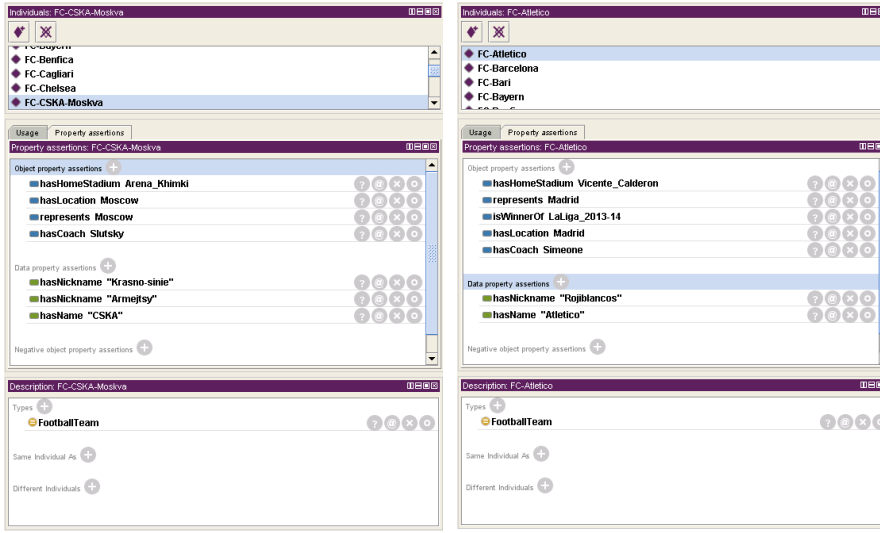


Fig. 7. Semantic structure of the sentence—answer



a b

Fig. 8. Background knowledge on the teams: a) CSKA; b) Atletico Madrid

6. Question Answering Based on Semantic Analysis

We have already mentioned that our semantic analyzer is applied to the task of answering natural language questions. To solve this task,

(1) natural language texts that presumably contain answers to the questions are processed by the semantic analyzer that builds a semantic structure for every sentence;
(2) the set of semantic structures generated by the semantic analyzer is converted into an OWL-document, which is uploaded into Protégé 4.3, a knowledge base management system;
(3) the semantic structure obtained for the question is converted into a SPARQL-query, which is then implemented at the SPARQL access point incorporated into Protégé 4.3.

Since the technique of semantic structure generation has been discussed in detail in Sections 3–5 above, we will focus here on the remaining two issues.

6.1. Generating an OWL-document for semantic structures

The language of semantic structure is based on the notions borrowed from the Ontology, which is represented in OWL. Respectively, the elements of semantic structures (individuals and their properties expressed with semantic relations) remain virtually unchanged when transformed into OWL elements. The individuals are assigned a special property **`belongToSent`**, which points to the sentence of the processed text that has served as basis for semantic structure generation. In this way, we arrange the semantic elements according to their first occurrence in the text.

Every OWL-document generated by the converter receives its own unique space of names (in IRI[3] format) which lists all individuals that are created in this document. The full name of such an individual includes a unique identifier which unambiguously determines the text where the individual was mentioned and even the sentence in which it first appeared. Due to this property we are able to match the facts with the texts from which they were extracted even if the respective OWL-document in which they were introduced is merged into some specialized repository of facts.

The following screenshot is a graphic Protégé 4.3 representation of an OWL-document fragment which was built by the semantic structure converter for the first sentence of text (1) from Section 4.

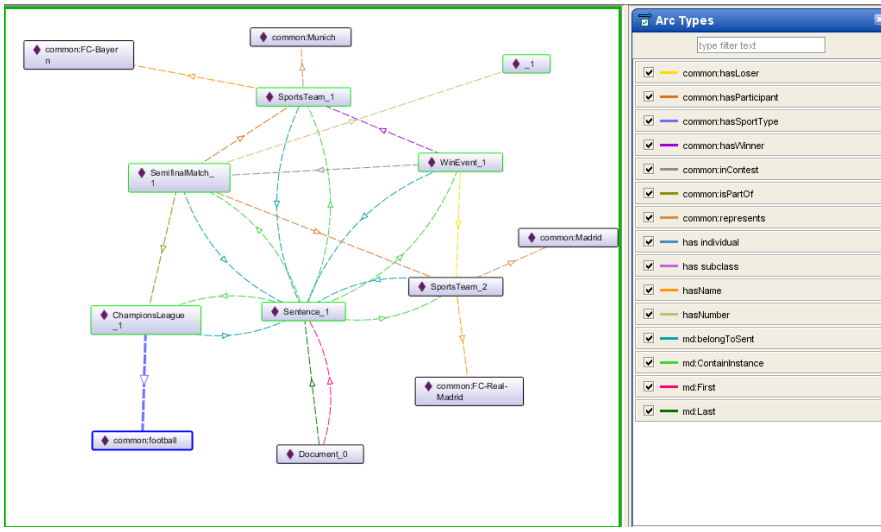


Fig. 9. An OWL-document representing a semantic structure

6.2. Generation of answers to natural language questions

The mechanism of answer generation used in our question answering system is based on the comparison of the semantic structure of the user question with structures present in the repository of individuals. The comparison consists in processing the repository with a SPARQL-query, generated from the semantic structure of the question Technically, SPARQL access point incorporated into Protege 4.3 is used for this purpose.

In the course of system development, we found that semantic structures for NL questions can be basically generated in the same way as the structures for declarative sentences from texts used to extract facts. One minor difference is the fact that we need to identify interrogative words, which is done by assigning these words with a special feature QUEST and forming a special name containing a question mark in the prefix. Such names instruct the SPARQL generator to determine whether

[3] Abbreviation of Internationalized Resource Identifier.

we have to do with a general or a special question. For general questions, SPARQL generates a so-called ASK-query which can be answered with a “yes” or “no”; whilst special questions generate a SELECT-query in which interrogative words are viewed as variables for which all possible values are selected and presented to the user.

Semantic structure elements that are not interrogative words are processed in ASK- and SELECT-queries in the same way. The processing consists in 1) substituting variables for individuals introduced in the question sentence; 2) using individuals specified in the Ontology in their original way: 3) transforming the semantic structure into a graph of triples, and 4) in referring individual variables to the respective classes.

The latter task may prove to be non-trivial if we need to check whether an individual indirectly belongs to a particular class. In this case, instead of the regular record of the form **`<variable> rdf:type <class>`**, where the variable and the class are connected with the relation rdf:type (to be an individual of class) we use the record like **`<variable> rdf:type/rdfs:subClassOf* <class>`**, which means that the variable and the class are linked by a string of relations, in which the first one is **`rdf:type`** followed by a zero (direct belonging to the class) or more relations **`rdfs:subClassOf (indirect belonging to the class)`**.

The following two figures exemplify queries generated for a special question (6) *Кого обыграла “Бавария”?* ’Who did Bavaria outplay?’ (Fig. 10) and a general question (7) *Мадридский «Реал» победил «Баварию»?* ‘Did Real Madrid defeat Bavaria’? (Fig. 11).

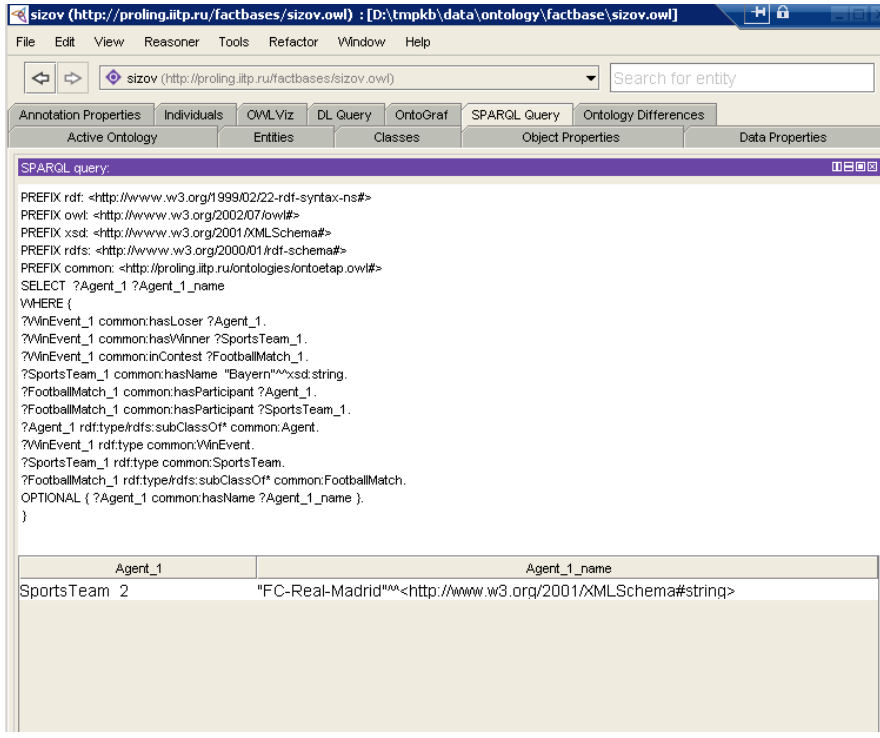


Fig. 10. A SPARQL query for a wh-question. The lower part shows the answer

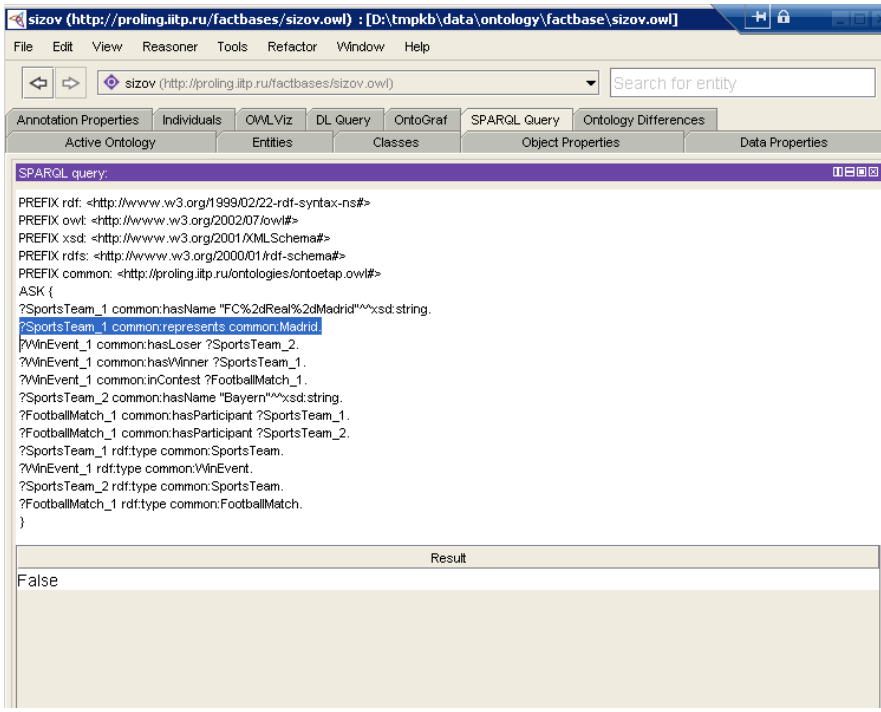


Fig. 11. A SPARQL query for a general question. The answer is negative

7. Future work and conclusion

We have presented a system of deep semantic analysis of Russian texts and shown how it can be used in a question answering system. The database used in our experiments is constructed on the basis of news texts, which are subject to a number of transformations. First, for every sentence a semantic representation is built. Semantic representations of sentences belonging to the same news message are merged into an integrated semantic structure using the coreference relation. The integrated semantic structure is converted into an OWL document. Natural language questions to the system are processed in a similar way: for any such question, a semantic structure is built, to be later converted into a SPARQL query to the database.

There are several important directions of work to be done. The specialized football ontology should be extended. On the one hand, many new instances (such as teams, players, stadiums) should be introduced and their properties described. On the other hand, many more concepts should be supplied with descriptions. In particular, we plan to develop a number of scripts to account for major complex events that occur during the match. One can wonder if it is at all feasible to compile and implement a reasonably complete list of scripts to be applied to match descriptions. In our opinion, this task is realizable, given that the total number of event types

in football is quite restricted, and each of them consists of a small number of elementary events (cf. our description of the scoring event above). Therefore, the task of identifying a complex event by its crucial components seems practicable.

A number of linguistic tasks need to be solved. In particular, the coreference resolution, which is the most sensitive component of the current system, should be made more reliable. As of today, we use a rather simplified rule: within a message, differently presented individuals belonging to one class are considered coreferential if there is no evidence to the contrary, like references to different participants of an event or different names of an object. This rule needs to be made more precise in order for system performance to improve.

Upon the extension of both the ontology and the linguistic data, a full-fledged evaluation of the system will be carried out.

References

1. *Akshay Java, Nirenburg S., McShane M., Finin T., English J., Anupam Joshi* (2007), Using a Natural Language Understanding System to Generate Semantic Web Content, International Journal on Semantic Web and Information Systems, 3(4), pp. 50–74.
2. *Allen J. F., Swift M., Beaumont W.* (2008), Deep Semantic Analysis of Text, Symposium on Semantics in Systems for Text Processing (STEP), volume 2008.
3. *Azmeh Z., Falleri J.-R., Huchard M., Tibermacine C.* (2011), Automatic Web Service Tagging Using Machine Learning and WordNet Synsets, Web Information Systems and Technologies. Lecture Notes in Business Information Processing. Vol. 75, pp. 46–59.
4. *Boguslavsky I. M.* (2011), Semantic Analysis based on linguistic and ontological resources, Proceedings of the 5th International Conference on Meaning-Text Theory (MTT'2011). Barcelona, September 8–9. pp. 25–36.
5. *Boguslavsky I. M., Dikonov V. G., Iomdin L. L., Timoshenko S. P.* (2013), Semantic representation for NL understanding, Computational Linguistics and Intellectual Technologies. International Conference Dialogue-2013 Proceedings, Issue 12 (19) in two volumes, Bekasovo, May, 29—June, 2. Moscow, RGGU Publishers, pp. 132–144.
6. *Bos J.* (2008) Wide-Coverage Semantic Analysis with Boxer, Semantics in Text Processing, STEP 2008 Conference Proceedings. W08–2222.
7. *Bos J.* (2011) A Survey of Computational Semantics: Representation, Inference and Knowledge in Wide-Coverage Text Understanding, Language and Linguistics Compass, Vol. 5, Issue 6, pp. 336–366.
8. *Bouayad-Agha N., Casamayor G., Mille S., Rospocher M., Serafini L., Wanner L.* (2012), From Ontology to NL Generation of Multilingual User- Oriented Environmental Reports, Proceedings of NLDB 2012: 17th International Conference on Applications of Natural Language Processing to Information Systems. Groningen.
9. *Clarke, J., Goldwasser D., Chang M. and Roth D.* (2010), Driving Semantic Parsing from the World's Response, Proceedings of the Fourteenth Conference on Computational Natural Language Learning (CoNLL-2010).

10. *Copestake A., Flickinger D., Pollard C. and Sag I.* (2006), Minimal recursion semantics: An introduction, Research on Language and Computation 3 (4), pp. 281–332.
11. *Coppola B. and Moschitti A.* (2010), A General Purpose FrameNet-based Shallow Semantic Parser, Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), Eds. Nicoletta Calzolari et al. Valletta, Malta: European Language Resources Association (ELRA).
12. *Dornescu I.* (2009), EQUAL: Encylopaedic Question Answering for Lists, Working notes for the CLEF 2009 Workshop. Corfu, Greece.
13. *Dukle K.* (2003), A Prototype Query-Answering Engine Using Semantic Reasoning, Master of Science Thesis, University of South Carolina. Manuscript.
14. *Ferrandez O., Spurk C., Kouylekov M., Dornescu I., Ferrandez S., Negri M., Izquierdo R., Tomas D., Orasan C., Neumann G., Magnini B., Vicedo J. L.* (2011), The QALL-ME Framework: A specifiable-domain multilingual Question Answering architecture, Web Semantics: Science, Services and Agents on the World Wide Web. Vol. 9, Issue 2, pp. 137–145.
15. *Ge R., Mooney R. J.* (2005), A Statistical Semantic Parser that Integrates Syntax and Semantics, Proceedings of the Ninth Conference on Computational Natural Language Learning. Ann Arbor, MI, pp. 9–16, June 2005.
16. *Iomdin L. L., Petrochenkov V. V., Sizov V. G., Leonid Tsinman L. L.* (2012), ETAP parser: state of the art, Computational Linguistics and Intellectual Technologies. International Conference (Dialog'2012). Moscow: RGGU Publishers, Issue 11(18). pp. 830–843. ISSN 2221-7932.
17. *Liang P., Jordan M., Klein D.* (2011), Learning Dependency-Based Compositional Semantics. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies — v. 1, p. 590–599.
18. *Mairal Usón R., Periñán-Pascual J. C.* (2009), The anatomy of the lexicon component within the framework of a conceptual knowledge base. Revista Espanola de Linguistica Aplicada 22, pp. 217–244.
19. *Moldovan D., Tatu M., Clark Ch.* (2010), Role of Semantics in Question Answering, Phillip C.-Y. Sheu, Heather Yu, C. V. Ramamoorthy, Arvind K. Joshi, Lotfi A. Zadeh (Eds.) Semantic Computing, pp. 373–420.
20. *Mueller E.* (2006), Common sense reasoning. Elsevier, Morgan Kaufmann Publishers.
21. *Nirenburg S., Raskin V.* (2004), Ontological Semantics. The MIT Press. Cambridge, Mass., London, England.
22. *Nirenburg S., McShane M.* (2012), Agents modeling agents: Incorporating ethics-related reasoning. Proceedings of the symposium Moral Cognition and Theory of Mind at the AISB/IACAP World Congress 2012, Birmingham, UK.
23. *Periñán-Pascual J. C., Arcas-Tunez F.* (2010a), Ontological Commitments in FungramKB. Procesamiento del Lenguaje Natural, p. 44.
24. *Periñán-Pascual J. C., Arcas-Tunez F.* (2010b), The architecture of unGramKB. Proceedings of ELRA Conference. Malta.
25. *Periñán-Pascual J. C., Mairal Usón R.* (2010), La Gramatica de COREL: un lenguaje de representation conceptual. Onomazein 21. Universidad de Chile.

26. *Poon H., Domingos P.* (2009), Unsupervised semantic parsing. Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing Volume 1 EMNLP 09 (p. 1).
27. *Raskin V., Taylor J.* (2010), Fuzzy Ontology for Natural Language. 29th International Conference of the North American Fuzzy Information Processing Society, Toronto, Canada, July 2010.
28. *Raskin V., Hempelmann C. F., Taylor J.* (2010) Application-guided Ontological Engineering. International Conference on Artificial Intelligence, Las Vegas, NE, July 2010.
29. *Riloff E.* (1999), Information extraction as a stepping stone toward story understanding, Ashwin Ram and Kenneth Moorman, editors, Understanding Language Understanding: Computational Models of Reading. MIT Press.
30. *Shi L. and Mihalcea R.* (2004), Open Text Semantic Parsing Using FrameNet and WordNet. Proceedings HLT-NAACL—Demonstrations '04 Demonstration Papers at HLT-NAACL 2004. pp. 19–22.
31. *Titov I., Klementiev A.* (2011), A Bayesian Model for Unsupervised Semantic Parsing. Learning Dependency-Based Compositional Semantics. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies — v. 1,USA, Oregon,.Portland. pp. 1445–1455.
32. *Tsinaraki C., Polydoros P., Kazasis F., and Christodoulakis S.* (2005), Ontology-based semantic indexing for mpeg-7 and tv-anytime audiovisual content. Multimedia Tools and Applications, Vol. 26, No. 3, pp. 299–325.

КВАНТИТАТИВНЫЕ МЕТОДЫ В ДИАХРОНИЧЕСКИХ КОРПУСНЫХ ИССЛЕДОВАНИЯХ: КОНСТРУКЦИИ С ПРЕДИКАТИВАМИ И ДАТИВНЫМ СУБЪЕКТОМ

Бонч-Осмоловская А. А. (abonch@gmail.com)

НИУ ВШЭ, Москва, Россия

Ключевые слова: дативные субъектные конструкции, предикативы, корпусная лингвистика, иерархическая кластеризация

QUANTITATIVE METHODS IN DIACHRONIC LINGUISTIC STUDIES: THE CASE OF RUSSIAN DATIVE SUBJECTS WITH PREDICATIVES

Anastasia Bonch-Osmolovskaya (abonch@gmail.com)

National Research University Higher School of Economics, Moscow Russia

The paper aims to demonstrate how quantitative corpus methods used in linguistics research may help to range different realizations of the same phenomenon: the use of dative subjects in predicative and adjective constructions. The core idea of the research is to study the distribution of dative subject constructions with predicative and adjective forms that potentially can be used in such constructions, i.e. aptitude of the construction for explication or omitting the dative subject. While usually the predicates are classified on the basis whether they can potentially be used with dative subject, I study the trends for explicit use of dative (or prepositional beneficiary arguments) among the “dative subject predicates”, and show that the frequency rates of real use of dative subjects can be very different with different predicates. Separate analysis of different morphological forms of the same dative subject lexeme (i.e. adjectives in full and short forms, comparative adjectives and predicatives) shows that they may also exhibit different strategies with explicit dative subjects. Finally data from the 18th and the 21st centuries is compared and hierarchical clustering is used to reveal some diachronic trends.

Key words: dative subject constructions, quantitative methods, corpus linguistics, Russian

1. Введение: теоретические подходы к анализу предикативных конструкций с дативным субъектом[1]

Конструкции типа *мне грустно* — с предикативами на -о и дативным субъектом обладают рядом противоречивых свойств, которые делают их исследование весьма непростой задачей. Во-первых, проблемой являются категориальные характеристики этих предикатов. Являются ли они лексически самостоятельными, относящимися к отдельной категории состояния (Щерба 1957, Виноградов 1947) и лишь омонимически совпадающими с соответствующими формам прилагательных и наречий? Или же они являются определенного вида деривацией, предикативными наречиями (Шведова 1980)? Или же они представляют собой несогласованную предикативную форму краткого прилагательного в среднем роде (дефолтный род согласования в русском языке) аналогично тому, как согласуются безличные глаголы (Babby 1974). Второй, безусловно связанной с пониманием категориальной природы предикативов на-о, проблемой является вопрос о продуктивности этих конструкций. Если принять парадигматический анализ Леонарда Бэбби и рассматривать предикативы и прилагательные как одну лексему, выступающую в случае предикативов в специальном мофросинтаксическом контексте, то оказывается, что далеко не все краткие формы прилагательных допускают употребление дативного субъекта. Более того, даже те из предикативов, которые могут употребляться безлично, лишь некоторые свободно присоединяют дативный субъект. Имеется достаточно большая «серая» зона предикативов, которые могут быть употреблены с дативным субъектом лишь при дополнительной контекстуальной поддержке, и при этом такие употребления кажутся сомнительными при изолированной интроспекции. Так, в предложении (1) употребление предикатива *знойно* с дативным субъектом является окказионализмом, который, благодаря контексту, легко интерпретируется, ср. изолированное употребление этой же конструкции в примере (2), которое выглядит существенно более проблематично.

(1) *ОК Жара описана хорошо — мне знойно стало и губы пересохли*
(Поиск по «Яндекс-блогам»)

(2) *? Мне стало* ***знойно***

При этом важно отметить, что в целом среди конструкций с дативным субъектом, конструкции с предикативами на -о весьма частотны: как было показано в (Бонч-Осмоловская 2003) предикативы на -о занимают третье по частотности место среди всех конструкций с дативным субъектом после экспериенциальных глаголов и модальных предикатов. Важно что, в отличие от экспериенциальных и модальных глаголов, списки которых исчисляются примерно

[1] В статье использованы результаты, полученные в ходе выполнения проекта (№14-01-0191), в рамках Программы «Научный фонд НИУ ВШЭ» в 2014-2015 гг.

двумя десятками, предикативов, требующих (или допускающих) дативное кодирование субъекта, на порядок больше. В работе (Циммерлинг 2010) сообщается о списке в 150–200 таких предикатов, и вопрос об их продуктивности остается открытым. Иначе говоря, в случае экспериенциальных глаголов мы можем однозначно сказать, что дативное кодирование субъекта есть словарное свойство предиката, поскольку у нас имеется внятный и законченный список дативных предикатов. С предикативами на-о само составление такого списка и определение того, насколько тот или иной предикат имеет валентность на дативный актант, является непростой исследовательской задачей.

В исследованиях последних лет было предложено несколько подходов к анализу семантических и синтаксических свойств предикативов на-о, которые так или иначе могут быть связаны с дативной субъектной конструкцией. Следует отметить, что в большинстве работ предикативы рассматривались как формы, парадигматически связанные с краткими и полными прилагательными (ср. подход Бэбби к определению категориальной природы предикативов, приведенный выше). В (Бонч-Осмоловская 2003, 2007) предикативы делятся на классы в зависимости от того, что происходит со вторым не дативным актантом. Такого актанта может вовсе не быть (*мне жарко*), или же он может быть выражен инфинитивом (*мне трудно дышать*), зависимой предикацией (*мне интересно, что ты ответишь*) или согласованной с кратким прилагательным номинативной именной группой (*мне понятен твой ответ*). Было показано, что имеются своего рода кластера предикативов, допускающих один или несколько вариантов оформления второго актанта. Так, в отдельную группу выделяются предикативы, встречающиеся либо без актанта вовсе, либо с инфинитивом. В другую группу попали предикативы, присоединяющие именную группу или предложение. Для предикатива *ясно* доминирующей стратегией является управление зависимым предложением. В работе была высказана гипотеза, что расширение списка предикативов происходит за счет наиболее продуктивного класса, который характеризуется отсутствием другого не дативного актанта.

В (Циммерлинг 2010) центральной идеей, классифицирующей предикативы на -о, является соотнесение дативной предикативной конструкции, которая в работе называется предикатом ситуативного признака LexP с адъективной формой — предикатом, соотнесенным с конкретным референтом. Постулируется три класса основ:

- амбивалентные основы, которые «образуются как предикаты, соотнесенные с конкретным референтом, так и предикаты ситуативного признака» *грустный — мне грустно*
- актантно — поляризованные основы, которые допускают только адъективные формы, а предикаты ситуативного признака (дативные субъектные конструкции с предикативом) имеют существенный сдвиг в семантике (ср. *жирный — мне жирно* в значении «слишком много»)
- ситуативно-поляризованные основы, которые не имеют на синхронном уровне соответствующей адъективной формой, или же кроме сдвига в значении имеется еще и сдвиг ударения (ср. слабый — мне слабó)

Циммерлинг вводит формат дескрипции свойств предикативов типа LexP в языке и сравнивает данные русского языка с данными литовского, исландского и чешского, в которых тоже имеются подобные конструкции. Формат включает в себя словообразовательную модель (источник, от которого образется форма), семантические характеристики дативного актанта, такие, например, как +одушевленность или — генеричность, объем класса LexP и продуктивность класса. Таким образом, фактически среди качественных прилагательных, т.е. таких, которые в принципе могут образовывать краткую и адвербиальные формы, выделяется подкласс амбивалентных основ, которые прежде всего и образуют предикативные конструкции с дативным субъектом.

Следующий существенный шаг в изучении свойств основ качественных прилагательных относительно их возможности употребления в дативных предикативных конструкциях был сделан в (Say 2014). В основе подхода также лежит парадигматическое понимание предикатива как непосредственно связанного с формами прилагательного с той же основой. В работе проводится корпусное исследование частотности употребления дативных субъектов с предикативами, согласованными краткими формами и полными формами прилагательных двенадцати лексем, про которые известно, что они могут использоваться в дативных предикативных конструкциях. В результате формулируются две импликации:

1. Если дативный субъект употребляется вместе с полной формой прилагательного, то он употребляется и с краткой формой. Обратное неверно.
2. Если краткая форма прилагательного может встречаться с дативным субъектом, то и соответствующий предикатив (несогласованная форма) тоже встречается с дативным субъектом. Обратное неверно.

Таким образом, в (Say 2014) формулируется иерархия дативных конструкций:

Предикативы > Краткие формы > Полные формы

Основы качественных прилагательных классифицируются на основании того, какую позицию они занимают в иерархии и, следовательно, образуют 4 класса:

- 1 класс не встречается с дативным субъектом (ср. например, **солнечный мне день, *мне солнечен день, *мне солнечно*)
- 2 класс употребляется с дативным субъектом только в конструкциях с предикативами (**грустная мне песня, *мне грустна песня, но* ^{ОК}*мне грустно*)
- Основы 3 класса встречаются с предикативами и краткими формами, но не с полными формами прилагательных (**тесный мне пиджак,* ^{ОК}*мне тесен пиджак,* ^{ОК}*мне тесно*)
- 4 класс допускает дативный субъект в любой из форм (^{ОК}*приятный мне человек,* ^{ОК}*мне он приятен,* ^{ОК}*мне приятно*)

Далее, анализируя семантические отношения между дативным аргументом и аргументной структурой соответствующих прилагательных, автор

показывает, что разные позиции на иерархии соотносятся с разными типами отношений. Датив может быть вообще никак не связан с аргументами прилагательного (тип 1:, напр. *холодно*), или может быть соотнесен с внутренним аргументом прилагательного (тип 2:, напр. *грустно*), или же непосредственно связан с дативным аргументом прилагательного (тип 3:, напр. *приятно*).

Изолированно каждый из таких типов может быть соотнесен с одним из теоретических подходов: тип 1 лучше всего описывается конструкционной моделью, именно поэтому этот тип является наиболее продуктивным и допускает образование контекстуально-обусловленных предложений с дативным субъектом, описывающих непосредственную реакцию экспериенцера (ср. пример(1)). Идея состоит в том, что конструкция PRED+NP_{DAT} имплицирует значение состояния экспериенцера, и сама возможность интерпретации лексических свойств предикатива как экспериенциального состояния обуславливается не столько его собственно семантикой, но контекстом описываемой ситуации в целом.

Для анализа предикативов второго типа наиболее близкой является деривационная интерпретация «дативных имперсональных конструкций» (см. Guiraud-Weber 1984). Фактически тот же самый деривационный процесс имеет место в деагентивных деривациях «*я не сплю* > *мне не спится*»

Наконец, третий тип наиболее четко укладывается парадигматическую концепцию Леонарда Бэбби: в этом случае предикативная конструкция с дативным субъектом есть ни что иное, как безличное употребление краткого прилагательного с сохранением его актантной структуры, в частности дативного аргумента.

Примирение трех подходов заключается в более абстрактном понимании конструкционной схемы, которая, по выражению автора, будучи «слепа к аргументной структуре соответствующего прилагательного», использует уже имеющийся датив для выражения экспериенцера, или же понижает внутренний аргумент до дативного слота, или же превносит его как дополнительный конструкционный элемент.

Еще одним важным замечанием, сделанным в выводах рассматриваемой работы, является методологическое наблюдение о зависимости теоретических моделей от привлекаемых данных. Как отмечает автор, расширение объема данных позволяет перейти от конкуренции моделей в сторону их контаминации.

Настоящая работа во многом развивает метод работы с данными, предложенный в (Say 2014) и основанный на обобщении корпусных наблюдений о поведении дативного субъектного аргумента в контексте соотнесенных форм предикатива (безличного употребления), краткой формы и полной формы прилагательного. При этом, в отличие от работы (Say 2014), в работе будут рассматриваться не все предикативы, а лишь такие, которые потенциально могут допускать конструкции с дативным субъектом во всех парадигматически связанных формах: обеих адъективных и собственно предикативной (описанные как тип 4 в (Say 2014)). Основная задача исследования видится в том, чтобы уйти от бинарного описания возможности/невозможности употребления дативного субъекта к более актуальной на сегодняшний день вероятностной оценке

«склонности» или «тяготения» (Stefanowitsch, Gries 2009) конструкции к эксплицитному дативному маркированию субъектного аргумента. Такой подход, во-первых, позволяет построить иерархию выраженной «дативной субъектности» внутри группы предикатов, которые в принципе допускают такие конструкции, а, во-вторых, дает возможность для диахронических исследований постепенных смещений конструкций в сторону одного или другого полюса.

Таким образом, фокус работы состоит в методологическом подходе получения новых данных о дативных предикативных конструкциях. Работа построена следующим образом: в части 2 будут выделены значимые параметры, на основании которых отбираются исследуемые данные, также будут описаны методы сбора и анализа данных, в части 3 будут рассмотрены результаты анализа данных с помощью метода иерархической кластеризации и предложена интерпретация полученных классов. Там же будут подведены предварительные итоги полученным наблюдениям и сформулированы задачи дальнейших этапов исследования.

2. Методы сбора и анализа данных

Основной концептуальной задачей исследования являлся уход от бинарного подхода к анализу предикативных конструкций (может или не может использоваться с дативом) и переход к анализу трендов — статистически значимых изменений, выражающихся в частотности употребления дативного субъекта. Неудовлетворительность бинарного подхода связана в первую очередь с самим характером данных, а именно наличием контекстуально-обусловленных конструкций, которые, как уже было сказано выше, не могут считаться приемлемыми при изолированном употреблении, однако встречаются в языке как спорадические новообразования.

Как показано в (Бонч-Осмоловская 2003), доля таких окказиональных конструкций в общей выборке предикативов с дативным субъектом весьма существенна. Кроме того, представляется чрезвычайно важным включить в поле рассмотрения диахронические данные, позволяющие проследить значимые изменения.

Исследование проводилось на материале Национального корпуса русского языка. Для исследования были выбраны десять частотных основ*: нужный, надобный*, *полезный, приятный*, *жаркий, слышный*, *понятный*, *плохой*, *худой* и *известный.* Важным свойством этих основ является то, что, во-первых, они достаточно распространены и хорошо представлены в корпусе. Во-вторых, за исключением основы *жаркий*, все они теоретически допускают употребление дативного субъекта с обеими адъективными и предикативной формами. В-третьих, они характеризуют разные стороны внутреннего состояния субъекта. Следует отметить, что наравне с дативным маркированием субъекта рассматривались и конструкции, в которых в этой же функции используется предложная группа с *для*, например, *полезный для меня*. Таким образом, исследовалось два вопроса: во-первых, оценивалась доля конструкций с дативным

субъектом для той или иной морфосинтаксической формы лексемы из списка, а во-вторых, были собраны данные о возможной конкуренции предложной и дативной конструкции для каждой такой формы.

Собственно материал исследования был представлен двумя выборками предложений, первая из которых была получена из текстов 18 века, а другая — из текстов 21 века[2]. Общий объем данных составил около 15 тысяч предложений. Такой значительный объем данных был необходим для надежных статистических подсчетов поведения 80 форм (по 4 формы каждой лексемы для каждого века).

Выбранный подход отличается от метода, используемого в Say 2014, в котором используется анализ 100 случайных употреблений для каждого из 12 предикатов. (О недостатках такого подхода см. Goldberg 2011). В настоящем исследовании каждая основа была представлена совокупностью вхождений конкретных форм в выборке текстов 18 или 21 века.

Полученные предложения были размечены вручную по ряду параметров. К каждому предложению была приписана базовая основа, была указана конкретная форма (полная или краткая согласованная адъективная, сравнительная степень[3] или же безличная предикативная), далее отмечалось наличие эксплицитно выраженного датива, а также наличие предложной группы с бенефициарным значением (в абсолютном большинстве случаев последние два параметра находились в дополнительной дистрибуции). Наконец, каждое предложение было также маркировано по принадлежности к выборке 18-го или же 21-го века.

Идея состояла в том, чтобы посмотреть на двух разных временных срезах, как соотносятся разные формы одной основы относительно дативного субъектного аргумента и предложного бенефактива. Поэтому основанием для сравнения стали подмножества данных, характеризующиеся пересечением трех общих параметров: общая лексема, один и тот же тип морфологической формы, общий диахронический показатель (век). Такого рода тройки параметров мы в дальнейшем будем называть типами. Для каждого типа (основа+форма+век) были подсчитано общее число вхождений, число вхождений с выраженным дативным субъектом, число вхождений с предложной бенефактивной группой. Абсолютные числа были нормализованы в проценты от общего числа вхождений. Всего мы получили 59 типов с данными, поскольку некоторые типы оказались вообще не представлены в корпусе (например, надобный в 21 веке), некоторые представлены слишком слабо для статистически надежных вычислений. Каждый из 59 типов характеризовался 3 числовыми значениями: количество предложений, в которых отсутствует экспериенциальный аргумент, количество предложений, в котором он выражен дативом,

---

[2] Для того, чтобы выборки были соразмерными, 21 век был ограничен только текстами, датирующимися 2003 годом.

[3] В целом формы со сравнительной степени имеют мало данных, поэтому, несмотря на то, что они будут присутствовать на рисунке 1, отражающем иерархическую кластеризации всех форм всех лексем обоих веков, эти формы пока не будет учитываться в результатах.

количество предложений, в котором он выражен бенефактивной предложной группой, в основном с помощью предлога *для*. В таблице 1 представлены для примера данные для типов, связанных с лексемой *нужный*. Колонка Total показывает общее количество вхождений этого типа в выборку, колонка None отражает количество предложений, в котором рассматриваемый тип был употреблен без выраженного дативного субъекта или бенефактивной предложной группы с предлогом *для*. Колонки Dat и PP соответственно отражают количество предложений с дативным субъектом или же с предложной бенефактивной группой. Следующие три колонки содержат те же данные, но выраженные в процентом соотношении. Даже из этого небольшого фрагмента таблицы видно и то, насколько по-разному ведут себя в отношении дативного субъектного аргумента разные морфологические формы одной и той же лексем, см., например, употребления предикатива *нужно* и прилагательного *нужный* в 18 веке с выраженным дативным субъектом — в первом случае количество таких предложений составляет 23% от общего количества вхождений типа, во втором случае только 5%. Очевидно также и то, что склонность к выражению дативного субъекта меняется со временем. Так, та же форма полного прилагательного *нужный* в 18 веке предпочитала конструкции с предложной группы для выражения субъектных отношений (*нужный для меня*) — таких конструкций обнаруживаем 15% от вхождений, а дативных только 5%. В 21 веке ситуация меняется противоположным образом, конструкций с выраженным дативным субъектом (*нужный мне*) становится уже 19% от общего количества вхождений типа, а конструкции с предлогом *для* становятся весьма редкими и составляют всего лишь 2%.

Таблица 1. Пример распределения разных конструкций внутри вхождения одного типа

TYPE	Total	None	Dat	PP	None%	Dat%	PP%
НУЖНЫЙ_18_ПРЕД	184	115	42	27	63	23	15
НУЖНЫЙ_18_КР_ПРИЛ	346	151	92	103	44	27	30
НУЖНЫЙ_18_ПОЛН_ПРИЛ	528	422	26	80	80	5	15
НУЖНЫЙ_21_ПРЕД	1443	1124	277	42	78	19	3
НУЖНЫЙ_21_КР_ПРИЛ	1255	405	752	98	32	60	8
НУЖНЫЙ_21_ПОЛН_ПРИЛ	330	258	64	8	78	19	2

Далее для анализа таблицы был применен метод иерархической кластеризации данных. С помощью этого метода наиболее близкие по поведению типы объединяются в группы, далее близкие группы объединяются в более общие группы. В результате все данные можно представить с помощью иерархического дерева (см. рисунок 1).

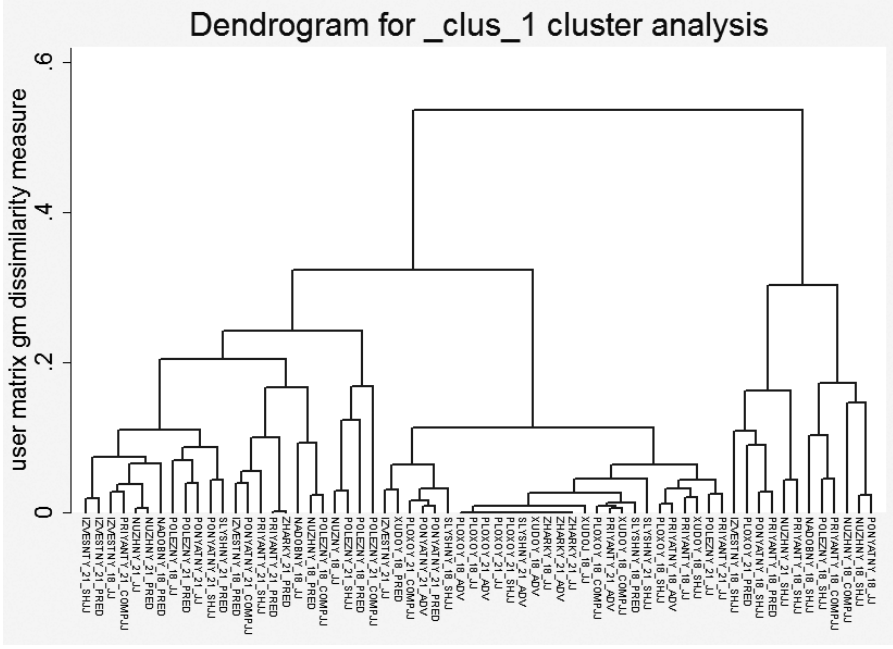


Рисунок 1. Дендрограмма близости типов основ относительно способов выражения экспериенциального элемента. Условные обозначения типов: JJ — полная форма прилагательного, SHJJ — краткая форма прилагательного, COMPJJ — сравнительная степень, PRED — предикатив, не управляющий номинативной ИГ и несогласованный с ней.

3. Результаты и обсуждение

Результаты кластеризации были интерпретированы вплоть до третьего уровня ветвления дендрограммы. Получившиеся в итоге кластеры были соотнесены с содержательными параметрами классов. Наполнение классов и параметры их выделения отражены в таблице 2. Ниже будут представлены краткие комментарии к анализу дендограммы.

Верхний уровень кластеризации разделяет все типы на те, у которых выраженный дативом экспериенциер или бенефактивная предложная группа встречается в более 40% случаев (экспериенциальный класс), и те типы, для которых частотность таких конструкций ниже (неэкспериенциальный класс). На втором уровне кластеризации в экспериенциальном классе выделяется подгруппа типов, для которых достаточно часто встречается и дативное выражение субъекта, и предложная группа с *для*. Отличительным признаком этой подгруппы является то, что у каждого из типов имеется больше 10% вхождений именно с предложной конструкцией. Такую подгруппу можно условно

называть предложным классом. Эта же подгруппа далее может быть разбита еще на две подгруппы — на собственно предложную, куда входит типы, у которых предложных конструкций больше, чем дативных, и смешанную — такую, в котором количество предложных дативных вхождений для типов представлено примерно поровну. Обратим внимание на то, что обе подгруппы предложного класса содержат в себе только типы, относящиеся к 18 веку. Иными словами, с помощью иерархической кластеризации выделяется, с одной стороны, группа предикатов, чье поведение кардинально изменилось в течение последующих двух веков, а с другой стороны, собственно класс поведения предикатов, который оказывается не представлен в 21 веке, по крайней мере, среди рассмотренных предикатов.

Другая подгруппа экспериенциального класса характеризуется использованием дативного маркирования для оформления экспериенцера, она может быть в свою очередь поделена на такую, в которой дативные конструкции доминируют и встречаются больше чем в половине всех вхождений типа (такой класс условно назовем супердативным), а также на группу, в которой дативные конструкции тоже составляют значительную часть вхождений типа, но все-таки чуть меньшую, чем в супердативной. Интересно, что в этой группе присутствуют и типы, относящиеся к 18 веку, и типы, относящиеся к 21 веку. Так же, как и предложный классы, дативный классы не очень многочисленен: и предложный класс, и дативный насчитывают по шесть типов.

Неэкспериенциальный класс на следующем уровне ветвления дендограммы подразделяется на группу, в которой выраженный экспериенциальный (или бенефактивный) аргумент все-таки встречается, и группу (так называемый смешанный класс), в которой экспериенциальное или бенефактивное отношение выражается совсем редко (так называемый слабый класс). Смешанный класс на следующем уровне дифференцирует типы по частоте встречаемости предложной группы. Заметим, что самым многочисленным классом среди всех кластеров является смешанный дативный класс — такой, в котором количество конструкций с дативным субъектом или предложной группой составляет менее 20%, но больше 10% от общего количества вхождений, и при этом дативные конструкции доминируют. В этот класс входят типы 18 и 21 века, всего к смешанному дативному классу относится 21 тип. При этом лишь типы *известно* и *нужно* представлены и вхождениями 18-го века, и вхождениями 21 века, иначе говоря, только эти два предиката сохраняют свое поведение в диахронии. Все остальные типы попадают в смешанный дативный тип только в 21 веке, или же как, например, *надобный* и *надобно* вовсе выходят из употребления. Ниже в таблице 2 представлены подробные параметры выделения каждого класса, и приведены все типы, которые попадают в тот или иной класс.

Таблица 2. Дифференциация типов предикатов с помощью иерархической кластеризации: основания для выделения классов и примеры вхождения типов в каждый класс

Класс 1	Экспериенциальный (Exp > 40%)	Предложный (PP > 20%)	Экспериенциальный предложный класс PP > DAT	**18:** нужен, нужнее, понятный
Класс 2	Экспериенциальный (Exp > 40%)	Предложный (PP > 10%)	Экспериенциальный смешанный PP|DAT	**18:** надобен, полезен, приятнее
Класс 3	Экспериенциальный (Exp > 40%)	Дативный класс (Dat > 20%)	Супердативный DAT > 55%	**18:** приятен **21:** нужен
Класс 4	Экспериенциальный (Exp > 40%)	Дативный класс (Dat > 20%)	Дативный 40% < DAT < 55%	**18:** известен, понятен, приятен **21:** плохо
Класс 5	Неэкспериенциальный (Exp < 40%)	Смешанный (20% < Exp < 10%)	Смешанный предложный PP > DAT	**18:** нужный, полезно **21:** полезен, полезнее
Класс 6	Неэкспериенциальный (Exp < 40%)	Смешанный (20% < Exp < 10%)	Смешанный дативный DAT > PP	**18:** известный, известно, надобный, надобно, нужно, полезный, полезнее, **21:** жарко, известен, известно, нужный, нужно, полезно, понятнее, понятный, понятен, приятнее, приятно, приятен, слышно, жарко
Класс 7	Неэкспериенциальный (Exp < 40%)	Слабый (Exp < 20%)	слабый DAT > 10%	**18:** слышен, худо **21:** известный, хуже
Класс 8	Неэкспериенциальный (Exp < 40%)	Слабый (Exp < 20%)	нулевой DAT < 10%	**18:** жаркий, плохой, плох, приятный, слышный, слышно, худой, худ, хуже, **21:** жаркий, плохой, плох, полезный, приятный, слышен

Получив сведения о диахронических изменениях частотности употребления дативной конструкции, мы можем проверить, насколько полученные нами данные соответствуют иерархии, сформулированной в Say 2014,

или — иначе — работает ли иерархия не на бинарном уровне категорического противопоставления (может/не может использовать дативный субъект), а на градуальном уровне противопоставления, который мы получаем благодаря делению предикатов на множество классов. Мы ожидаем, что если количество вхождений с выраженным дативным субъектом увеличивается у полной формы, то тогда, это количество будет увеличиваться у краткой формы и у предикатива. Противоречить иерархии будет ситуация, когда количество конструкций с выраженным дативным субъектом у более низкой позиции (предикатива) понижается, а у более высокой позиции (напр. краткое прилагательное) увеличивается. Представим данные таблицы 2 в виде отдельной таблицы, отражающей изменения каждой из форм: курсивом будет отображаться употребление 18 века, а подчеркиванием употребление 21 века. При этом заметим, что иерархия, предложенная в Say 2014, рассматривала исключительно дативные аргументы, а не бенефактивные предложные группы. Таким образом, мы также можем проверить, работает ли иерархия относительно употребления предложных групп с предлогом *для*. В таблице 3 представлена динамика изменений принадлежности к классам типов, относящихся к одной лексеме

Таблица 3. Диахронические смещения по принадлежности к классам морфологических типов лексем. Курсивом выделены типы, относящиеся к 18-му веку, а подчеркиванием к 21-му. В таблице приведены краткие обозначения классов, полные обозначения и обоснования их выделения см. в таблице 2.

экспериенциальный				**неэкспериенциальный**			
предложный		**дативный**		**смешанный**		**слабый**	
предлож	**смеш**	**супер-датив**	**датив**	**пред-лож**	**датив**	**слабый**	**нулевой**
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
понятный			*понятен*		понятный понятен	понятно	
нужен	нужен			*нужный*	нужный нужно		
			известен		*известный* известен *известно* известно	известный	
	полезен			полезен *полезно*	*полезный* полезно		полезный
			приятен *приятно*		приятен приятно		*приятный* приятный
					жарко		*жаркий* жаркий

экспериенциальный				**неэкспериенциальный**			
предложный		**дативный**		**смешанный**		**слабый**	
предлож	**смеш**	**супер-датив**	**датив**	**предлож**	**датив**	**слабый**	**нулевой**
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
			плохо				*плохой* плохой *плох* плох
					слышно		*слышен* слышен *слышно*

Столбики с классами в таблице упорядочены в последовательности уменьшения доли дативных субъектных конструкций. Следуя предсказаниям иерархии, предложенной в Say 2014, мы ожидаем, что любая краткая форма прилагательного будет расположена в таблице не правее полной формы прилагательного той же лексемы, а предикатив той же лексемы будет занимать позицию не правее краткой формы прилагательного. Можно заметить, что в целом это предсказание соблюдается. Однако необходимо сделать несколько оговорок.

Во-первых, особым образом ведет себя предикатив *понятно* в 21 веке: дативные конструкции с этим предикативом встречаются реже, чем с полной и краткой формой прилагательного. У нас слишком мало данных об предикативном безличном употреблении соответствующей лексемы в 18 веке, поэтому в итоговую выборку вхождения *понятно* в 18 веке не вошли. Иными словами, мы не можем посмотреть динамику изменения поведения этого предикатива, но можем только зафиксировать текущее положение вещей, нарушающее иерархию. Возможное объяснение тут состоит в распространении употребления предикатива *понятно*, как вводного слова, не имеющего собственной аргументной структуры:

(3) *А в «Терре» запросто, без всяких этих штучек, так что она мне как раз подходит. Тебе-то,* ***понятно****, там не нравится. Тебе бы как раз в эти ваши клубы ходить…* [Анна Берсенева. Полет над разлукой (2003–2005), ruscorpora.ru]

(4) *Плата за отопление, воду, канализацию и электричество в эту сумму,* ***понятно****, не входит: все это жильцы получают не от своего товарищества, а от городских служб — им и платят.* [Владимир Абгафоров. Плата за проживание (2003) // «Мир & Дом. City», 2003.04.15, ruscorpora.ru]

Далее, поведение, противоречащее иерархии, обнаруживает предикатив *нужно*, который расположен правее краткого прилагательного *нужен*. Опять же у нас недостаточно данных для понимания того, к какому классу относился бы *нужно* в 18 веке, и мы не можем проследить динамические изменения в поведения этой формы. Заметим, впрочем, что семантически близкий предикатив *надобно,* используемый в 18 веке в том же контексте, а потом вытесненный *нужно,*

относится к тому же самому, что и *нужно*, популярному классу неэкспериенциальных дативных типов.

Наконец, краткое прилагательное *полезен* показывает в 21 веке существенную относительно других форм этой лексемы долю конструкций с предложной группой. При этом, иерархия выражения дативного субъекта нарушается и в 18 веке. Предикатив *полезно* в 18 веке имеет меньшую долю дативных конструкций по сравнению с соответствующим кратким прилагательным, у *полезно* количество таких вхождений равно 16%, а у предикатива *полезен* достигает 32% от общего количества вхождений этого типа. В 21 веке, *полезно* встречается с дативным субъектом в 11% случаев, а *полезен* лишь в 7%, однако в 14% случаев с предложной бенефактивной группой. Таким образом, уменьшение дативных субъектных конструкций отчасти компенсируется распространением бенефактивно-предложного оформления, но только в контексте кратких прилагательных.

Отметим также, что сдвиг большого количества типов в сторону смешанного дативного класса, и тем более существенное расширение использования предикативов, таких как *слышно* или *жарко*, указывает на то, что мы имеем дело с продуктивной лексико-грамматической конструкцией. Следует уточнить, что под конструкцией, в данном случае, имеет смысл понимать не схематическое описание Dat+Pred, но скорее совокупность распределения выраженного и не выраженного датива.

Итак, в работе был опробован метод иерархической кластеризации на диахронических данных. Такой метод ориентирован на изучение языкового употребления и позволяет фиксировать изменения в этом употреблении, которые в дальнейшем ведут к более существенным лексическим и грамматическим сдвигам. Анализ получившихся классов иерархической кластеризации данных по экспериенциальному маркированию субъекта (дативного и предложного) показал, с одной стороны, очевидные тенденции к унификации использования дативной конструкции, а с другой стороны, выделил случаи, которые по каким-то причинам не укладываются в направление общего сдвига. Накопление и анализ более значительного количества данных позволит в дальнейшем генерализовать предложенные выводы.

Литература

1. *Бонч-Осмоловская А. А.* Конструкции с дативным субъектом в русском языке, Диссертация на соискание ученой степени кандидата фил. наук, Москва, 2003.
2. *Бонч-Осмоловская А. А.* Конструкции с дативным субъектом с предикативами на -о/-е, Киселева, Плунгян, Татевосов (ред), Корпусные исследования по русской грамматике, Москва, 2009.
3. *Виноградов, В. В.* Русский язык (грамматическое учение о слове). Москва — Ленинград, 1947.

4. *Циммерлинг А. В.* Именные предикативы и дативные предложения в европейских языках // Компьютерная лингвистика и интеллектуальные технологии, По материалам ежегодной Международной конференции Диалог. — 2010.
5. *Шведова Н. Ю.* (ред.) Русская грамматика, Москва, 1980.
6. *Щерба Л. В.* О частях речи в русском языке // Избранные работы по русскому языку. — 1957. — С. 63–84.

References

1. *Babby, L. H.* 1974. Towards a formal theory of "parts of speech". In Slavic transformational syntax. Brecht, R. & Chvany, C. (Eds). Ann Arbor: Univ. of Michigan. 151–181.
2. *Bonch-Osmolovskaya A. A.* Dative subject constructions in Russian (Konstruktsii s dativnym subyektom v russkom yazyke), Dissertation, Moscow, 2003.
3. *Bonch-Osmolovskaya A. A.* 2009 Dative subject constructions with predicatives on -o/-e (Dativnye subyektnye konstruktsii s predikativami na -o/-e), Kiseleva, Plungyan, Rakhilina, Tatevosov (eds) Corpus studies in Russian grammar (Korpusnye issledovaniya po russkoy grammatike), Moscow, 2009.
4. *Goldberg A. E.* Corpus evidence of the viability of statistical preemption // Cognitive Linguistics. — 2011. — Т. 22. — № 1. — С. 131–153.
5. *Say S.* On the nature of dative arguments in Russian constructions with "predicatives" // Current Studies in Slavic Linguistics [Studies in Language Companion Series, 146]. Amsterdam: John Benjamins. — 2013. — С. 225–245.
6. *Shcherba L. V.* On parts of speech in Russian, Selected works on Russian language, Leningrad, 1957, pp. 63–84.
7. *Shvedova N. Yu.* (ed) Russian Grammar, Moscow, 1980.
8. *Stefanowitsch A., Gries S. T.* Collostructions: Investigating the interaction of words and constructions // International journal of corpus linguistics. — 2003. — Т. 8. — № 2. — С. 209–243.
9. *Vinogradov V. V.* The Russian Language (grammatical theory about word) (Russkiy Yazyk (grammaticheskoe ucheniye o slove)), Moscow-Leningrad, 1947.
10. *Zimmerling A. V.* Nominal predicatives and dative clauses in European languages (Imennye predicativy I dativnye predlozheniya v evropejskikh yazykakh) Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2010" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2010"], Bekasovo.

ГРАММАТИЧЕСКИЙ СТАТУС ЧЕРЕДОВАНИЕ НАЧАЛЬНОГО [Н]/[J] В ЛИЧНЫХ МЕСТОИМЕНИЯХ ТРЕТЬЕГО ЛИЦА

Даниэль М. А. (misha.daniel@gmail.com)

НИУ Высшая школа экономики
Московский государственный университет
им. Ломоносова
Университет Хельсинки
Москва, Россия

Ключевые слова: морфология, сандхи, контролер, вариативность, русский язык, местоимения, предлог, падеж, Национальный корпус русского языка

STEM INITIAL ALTERNATION IN RUSSIAN THIRD PERSON PRONOUNS: VARIATION IN GRAMMAR

Daniel M. A. (misha.daniel@gmail.com)

National Research University Higher School of Economics[1]
Moscow State University
University of Helsinki
Moscow, Russia

The paper discusses the present stage of the evolution of the initial [n]/[j] stem alternation in Russian third person pronouns. After providing a short overview of the origins of the forms, I focus on their category status, discuss Zalizniak's 'adpositionality' in some detail, and then proceed to considering the cases where the 'n'-forms are induced by a distant 'controller'. I will show that the fact that the 'n'-forms are essentially variants is better accounted for by the notion of 'trigger' of a morphological variant. To my eyes, this opens

[1] The study was implemented in the framework of the Basic Research Program at the National Research University Higher School of Economics (HSE) in 2015. Unless otherwise indicated, all examples come from Russian National Corpus (www.ruscorpora.ru). Although most examples come from corpus, I hardly use corpus statistics, partly because the key evidence, discussed in the concluding part of the paper, is numerically very weak in RNC and requires further research across blogs and similar linguistic data.

ways to a better understanding of the observed evidence than that using the conventional notion of morphosyntactic controller, on the one hand—and certainly than explaining them in (morpho)phonological terms. In the end, I will briefly argue that, in a sense, the evolution of the alternation is similar to degrammaticalization, showing a movement from a morphophonologically conditioned external sandhi to a morphosyntactic category similar to government.

Keywords: morphology, sandhi, pronouns, Russian, preposition, case, Russian National Corpus

1. Introduction

In modern standard Russian, there is a contextually conditioned alternation of the initial [j] (in some contexts, including pronouns, realized as Ø when followed by an [i]) with [n] in the oblique cases of third person personal pronouns of all genders and numbers. Cf. the dative forms: *ему* ~ *нему*, *ей* ~ *ней*, *им* ~ *ним*. The 'n'-forms are obligatory with what is traditionally considered to be primary prepositions (cf. 1; sometimes the alternation is used as a diagnostic for primary prepositions) but may be optional to impossible with other prepositions[2]. It also appears, optionally rather than obligatorily (2, 3), after comparative forms of adjectives and adverbs. On the other hand, the form is fully ungrammatical when the pronoun depends on the verb, as in (4).

(1) *Не ладил* ***с ним*** (*им) *один Григории Иванович Муромский, ближайший его сосед.*
[А. С. Пушкин. Барышня-крестьянка (1830)]

(2) *Она решительно опустилась на бревно сзади солдата и* ***выше его*** *и негромко, мягко, но строго заговорила…*
[Максим Горький. Солдаты (1906–1908)]

(3) *Над дверью сеновала, для отвода болезней от лошадей, был прибит гвоздями скелет птицы, на коньке крыши торчал чисто вымытый дождями рогатый череп козла,* ***выше него*** *неустанно качались голые вершины деревьев.*
[Максим Горький. Городок Окуров (1909)]

[2] Cf. (an almost) obligatory use of [n]-forms with *без* or *от* or *перед* but only peripherally with *вослед* (no occurrences in Russian National Corpus, few occurrences resulting from a Google query—it is truly hard to find examples that are completely un-googleable, for any of the secondary prepositions). The list of prepositions requiring 'n'-forms is discussed e.g. in (Itkin 267ff); but some judgments of the author are not uncontroversial. Thus, it seems that with *подле* they are not absolutely obligatory (less obligatory than e.g. with *без*). A corpus based study of 'n'-forms with different (groups of) prepositions is required, which however lies outside the qualitative aims of this paper.

(4) *В ярости тот схватил профессора за горло,* ***повалил его*** (*него) *на пол и задушил бы, если бы Елена Карловна не выстрелила…*
[Леонид Юзефович. Князь ветра (2001)]

This paper primarily considers the category status of the stem initial alternation. I will first briefly introduce the origins of this alternation, then will discuss the approach to the phenomenon adopted in (Zalizniak 1967), and will turn to the cases of 'distant control' of the alternation and their relevance for the interpretation of the phenomenon.

2. Origins and the present status of the form

Due to the frequent use of primary prepositions, frequency of some of the 'n'-forms in the corpus is comparable to that of the 'j'-forms. Historically, however, they are clearly secondary and originate from a morphological (or even morphophonological) reanalysis.

In Old Church Slavonic, several prepositions took a prothetic final [n] before a vocalic anlaut of the next wordform (Shakhmatov 1957: 162; Lunt 2001: 63). In pronouns, [jV] anlaut counted as vocalic (Polivanova 2014) and required the prothetic [n] of the preposition. This consonant, still present in some lexicalizations (e.g. *в(н)-утрь*, cf. *утроба*), has been consistently preserved only in combinations of prepositions with third person pronouns. As a result, it was reinterpreted as the component of the pronoun if preceded by a preposition, substituting the initial [j], thus turning into a contextually conditioned alternation. There then was an expansion of 'n'-forms to other contexts, including other, more recently grammaticalized prepositions (e.g. *напротив него*) and, ultimately, to comparative constructions lacking prepositions at all, as in (3).

This paper focuses on the category status of the alternation. One approach to paradigmatic analysis is suggested in (Zalizniak 1967). The forms in [n] are referred to as non-standard adpositional series (*нестандартная припредложная форма*; not to be confused with the prepositional case—*предложный падеж*) that result from the standard forms by adding "н" (orthographically speaking). He argues that the alternation may be considered a separate pronominal inflectional category cross-cutting the category of case and number (in pronouns). While most contexts require 'j'-forms, contexts with prepositions as a head require adpositional forms (which amounts to a kind of government, but category rather than lexically based). Zalizniak also explains apparent gaps in the inflectional paradigms of 'adpositionality' (*припредложности*). Out of all cases, the nominative does not have the adpositional form, and the prepositional case does not have the non-adpositional one. This is readily explained by the fact that the nominative is not used adpositionally, while the prepositional case is never use otherwise; the blank cases in the paradigm are excluded by structural reasons rather than arbitrarily. Under this view, prepositions are syntactic controllers of the inflectional category of adpositionality.

After having introduced the new inflectional category, Zalizniak however suggests that these forms are more conveniently treated as *variants* of the non-adpositional, or standard, forms. This line of argument is not unlike his treatment of the

second genitive (alias partitive) and the second prepositional (alias locative) cases. For these forms, too, he first considers a strictly structural, paradigmatic solution, but then explicitly re-considers this solution because it leads to over-sophisticated models of inflection. Such shift from a purely structural to what may be called variance approach to the forms both greatly simplifies the architecture of the whole system and makes the resulting model more intuitive and closer to the traditional description of the categories in question. Importantly, it is based on considering the adpositional form as a contextually determined variants of the realization of a category (e.g. **adpositional variant** of the dative value of the case category) rather than a separate value of a different category (**adpositional** singular dative as opposed to **non-adpositional** singular dative).

In this approach the issue of the prepositional case becomes somewhat more problematic. The adpositional form of the prepositional case (as in *в нём*) is then a variant of a form that does not exist at all—a strange status for a variant. Zalizniak circumvents this problem by suggesting an abstract non-adpositional form (**ём*) of which the adpositional form is a variant. This, however, seems to contradict the intuitive understanding of variation.

While treating all paradigmatic questions in fine detail, considering the conditions of use of the 'n'-form lies without the scope of Zalizniak's research. He quotes contexts where the forms tend to be obligatory rather than optional. He indicates passim that suggesting that the forms are used after prepositions is a simplification, but does not elaborate on that. Indeed, examples above show that the form may also be used with comparative forms of adjectives and adverbs. Although I consider the term adpositional form as a very convenient label, I prefer to use a different one that would be neutral to the apparently false assumption that the 'n'-forms are only used after prepositions.

3. Contact position

One important fact about the use of the form seems so obvious that it sometimes fails to be mentioned. In all of the examples above, the adpositional form of the pronoun immediately follows the wordform that requires it, be it a preposition or a comparative form. Although very natural from the historical perspective (after all, the adpositional forms have developed out of reanalysis of the preposition-and-pronoun complex and thus require adjacency) this fact requires some theoretical consideration.

Indeed, positing a separate pronominal category of adpositionality does not explain why the preposition, the 'controller' of the adpositional category, should be adjacent to the adpositional form, control being a syntactic rather than a linear notion. One answer to this could be that combinations of pronouns with prepositions naturally require adjacency. What is it that could go in between the pronoun and the preposition in a prepositional phrase? To my eyes, however, this requirement of adjacency (to the extent that it indeed holds—see below Section 4) makes the opposition between 'n'- and 'j'-forms look more like a morphophonologically determined external sandhi than a morphosyntactic category.

On different grounds, Bulygina (1977: 198) argues against adpositionality as a nominal inflectional category. Consider the use of 'j'-forms in coordination, as in *передо мной и ею* vs. *перед нею*. Morphosyntactic interpretation of the alternation as an inflectional category would lead to analyzing the two forms as different morphosyntactic values in the same syntactic environment. Bulygina suggests that the choice of the form is governed by phonological rules rather than controlled by prepositions. As we will see below, her points are controversial, but at the same time they provide important insights.

Let us consider this issue in some detail.

The adjacency requirement indirectly supports the morphophonological approach to the alternation. Considering the conditions of the alternation is however ambiguous. On the one hand, the class of contexts where the 'n'-form appears seems to be defined in syntactic rather lexical or morphological terms. Indeed, whenever the form is used, its presence is licensed either by a preposition or by a comparative form. On the other hand, there is a clear variation in the texts in that some prepositions obligatorily require the 'n'-form while some other prepositions may occur with the 'n'- or 'j'-forms (in RNC, one of the factors is when the text was written), and comparative forms mostly combine with 'j'-forms but a few may also take 'n'-forms. All this should of course be subject to quantitative analysis, but it is clear that, while the domain of use are limited by syntactic conditions, within this domain the alternation seems to be partly governed by lexical properties of the controllers.

The following well-known fact seems to be important. Adjacency alone does not suffice for alternation: if a pronoun follows a preposition but its syntactic head is elsewhere, the alternation does not happen:

(5) *Нелли Сергеевна, утвердясь* ***в*** [***ей*** *по всем законам и правилам принадлежащей жизни], скараулила момент, когда свекровь была во дворе, Марина в дровянике, и громко, ровно бы глухим, заявила <…>*
[Виктор Астафьев. Пролетный гусь (2000)]

In this example, the pronoun immediately follows the preposition but may not alternate with the 'n'-form. The reason is that it depends not on the preposition but on the verb (in this case, participle), so that the preposition and the pronoun are separated by NP boundary; see discussion in (Itkin 2007: 268).

This is however not enough to discard morphophonological interpretation altogether. It is known that external sandhi may be sensitive to the syntactic structure. Thus, liaisons in French are e.g. more frequent within NP than on its boundary, and initial lenition in Nivkh is only possible within certain syntactic constituents. Such alternations may be limited to more closely-knit units; the immediate cause of having or not having liaison may be prosodic unity. Unless these prosodic conditions may be consistently generalized in (morpho)syntactic terms[3], they remain a gray area between phonology and syntax, so that belonging to the same syntactic group may

[3] In fact, Nivkh initial 'lenition' is often considered to be evidence for incorporation; see discussion in (Mattissen 2003).

be considered to be a more distant cause, and the phonetic boundaries of the phrase the more immediate one.

As Section 4 shows, the case of the Russian 'n'-forms is different in that they are syntactically, not phonologically conditioned in the first place; sensitive to what they depend on (prepositions of comparatives) rather than what immediately precedes them. It would be easier to explain the absence of alternation in (5) by saying that the pronoun does not depend on the preposition. Nevertheless, one notes a certain similarity between the stem initial alternation in Russian pronouns and lenition in Celtic or Nivkh in terms of morphophonological technique involved.

4. Distant position

Most importantly, the assumption above that 'n'-forms are limited to the cases where they are immediately preceded by a preposition or a comparative is falsified by empirical evidence from texts. Some of such cases, which blatantly contradict Bulygina's (1977) claim about phonological nature of the conditions of the alternation, were reviewed in (Itkin 2007: 268) and discussed in (Yemtseva 2011). There are two such syntactic construction attested so far. In the first construction, the pronoun and the controller are separated by the agreeing form of the pronoun *весь* or *сам*. In the second, the pronoun is the second constituent in a co-ordinative construction introduced by the controller of the alternation:

(6) *Не в последнюю очередь потому, что хорошо знал людей из самого близкого окружения Брежнева и относился почти* ***ко всем ним*** *с глубокой антипатией.* [Георгий Арбатов. Человек Системы (2002)]

(7) *Такие случаи собственного бессилия* ***на самого него*** *нагоняли какую-то подавлявшую тоску, и он понимал состояние Нюрочки.* [Д. Н. Мамин-Сибиряк. Три конца (1890)]

(8) *Стас, стараясь держаться за «Москвичом» не впритык, а оставляя* ***между собой и ним*** *одну или две машины, ехал и размышлял о Нонне.* [Н. Леонов, А. Макеев. Эхо дефолта (2000–2004)]

A corpus query gives a very limited range of examples for both construction: 35 and 4 for the first (with *весь* and *сам* respectively) and 31 for the second. A more extensive study based on internet usage is required to make conclusions about the diachrony of the alternation with a distant controller. Below I will only venture a preliminary qualitative discussion of the impact of this evidence on our view on the nature of the alternation. Several quantitative observations, however, are obvious even on such small a scale. First of all, in the corpus (but not in the internet usage) the coordinative construction, as in (8), occurs exclusively with the preposition *между* (or its rare variant *меж*), while the contexts as in (6) and (7) do not seem to show obvious prepositional preferences. Second, the contexts have very different corpus

histories. The construction with *весь* first occurs in the end of the 20th century; all other occurrences fall between 2000 and the present. The construction with *сам* and *между*, however, are evenly distributed along the whole timeline of the corpus, first occurring in a text dated 1821 and last in 2010 for *между* and 1768 and 2005 for *сам* (in principle, this may be an indication that they are fading away, because of the increase in the size of the corpus with time and especially in the 2000s to the present).

What is this that makes the preposition *между*, of all prepositions, a likely distant controller of the alternation in coordination, and an early distant controller in the corpus? In a co-ordinative construction controlled by *между*, the two co-ordinands seem to be in a closer syntactic relation with the preposition than is the case with other prepositions. *Между* is, semantically, a non-unitary predicate whose arguments may be filled either by multiple referents of a plural NP (*между деревьями*) or co-ordinands (*между морем и небом*); see (Valova 2012 for results of a corpus study of *между* and its plural arguments). Most other prepositions are one-place predicates that take the coordinated NP as a whole. *Среди*, another semantically non-unitary preposition is however not attested as a distant controller (probably because, unlike *между*, it clearly prefers single plural dependent; note also that it is more than two times less frequent in the corpus).

Let us now briefly review the constructions with *весь*[4]. In the absence of robust quantitative data, I will limit myself to a qualitative interpretation. They may fall into three different groups, exemplified below with contexts involving the preposition *к*(*о*):

(9) *А панегирических суждений не привожу за их избыточную восклицательность и единообразие и потому, что* ***ко всем им*** *присоединяюсь, конечно.*
[Наталья Шмелькова. Последние дни Венедикта Ерофеева (2002)]

(10) *Не в последнюю очередь потому, что хорошо знал людей из самого близкого окружения Брежнева и относился почти* ***ко всем ним*** *с глубокой антипатией.*
[Георгий Арбатов. Человек Системы (2002)]

(11) *Поэтому* ***к ним ко всем*** *отношусь с почтением и уважением.*
[Владимир Губарев, Иосиф Фридляндер. Академик Иосиф Фридляндер: «трижды могли посадить...» // «Наука и жизнь», 2006]

All these constructions occur in the corpus; in the case of the preposition *к*, each occurs only once. Following some of the previous studies, Itkin (2007) reviews the alternatives and says that 'n'-forms are required after preposition followed by *весь* just in the same way as they are required immediately following the preposition, thus implicitly ruling cases like (9) out. This, however, is not at all supported even by the meager corpus evidence I presently have; in fact, with different prepositions 'j'-forms seem to dominate.

---

[4] I do not consider contexts with *сам* because of their low frequency in the data; but in principle my conclusions should be equally applicable to this construction, too.

The conclusions that I draw from examples (9) to (11) are as follows.

(a) distant control of alternation with prepositions is possible;
(b) unlike adjacent control with the same preposition, it is not obligatory; and
(c) a third alternative is also used (probably as a conflict-resolving alternative—cf. Itkin 2007).

In terms of optionality of alternation, the non-adjacency is thus added as a new dimension to the other dimensions of variation established above: the distant *κ* is a weaker *trigger* of alternation than the contact one, in the same way as comparative forms are weaker triggers than prepositions in the contact position, and some prepositions are weaker triggers than others.

Note that a true morphosyntactic controller may not depend on the linear distance but only on its syntactic relation to the target. I suggest that there is more than convenience considerations to the 'variance' solution that Zalizniak (1967) follows in the end, at least from the synchronic point of view. Considering 'n'-forms as variants of case forms rather than realization of a value of a separate category not only corresponds to the intuition but also opens a way to the explanation of non-adjacency effects.

Indeed, we have three types of contexts: those that require 'n'-forms, those that favor them and those that allow them (as well as, trivially, those that disallow them). To account for the morphosyntactic variation, both across types of contexts and across texts, I suggest to substitute the notion of controller that *requires* a certain value of a *morphosyntactic category* with the notion of a trigger that *allows* to *favors* to *requires* a *non-standard realization* of a form. Prepositions are, on the whole, stronger triggers of stem initial alternation than comparatives. There seem to be no comparatives that require alternation even when in adjacent position. Most adjacent preposition, on the other hand, are actually so strong a triggers that they *require* alternation and, in this position, are indistinguishable from true controllers. Distant prepositions, on the other hand, allow but do not require it. Whether they favor it or not is a question that requires further study on a different, fuller array of data.

In this approach, based on the notion of morphosyntactic variant, it seems natural that adjacency is a factor. If a variant is contextually associated with a certain trigger, the closer the trigger is, the more probable is the use of the variant. It seems that the same logic explains distributional facts about other variant forms in Russian, such as the second genitive and the second prepositional case[5].

5. Conclusion

The data above suggests a view on the evolution of the morphosyntactic status of the initial stem alternation in Russian third person pronouns. Starting from a clear

[5] Forms that, too, are first classified as separate categories by Zaliznak, but then merged with the standard forms 'for the sake of convenience'; similarly to the category of adpositionality, the first approach is criticized in (Bulygina 1977).

case of external sandhi (prothetic [n] before the initial vowel of the next wordform) it then changes its locus through reanalysis (from stem final prothesis in prepositions to stem initial alternation in pronouns) and thus narrows its scope in terms of the lexical involvement (to the pronominal forms) but ultimately expands it in terms of syntactic conditions (to other prepositions). At this stage, it becomes similar to a morphosyntactic category (cf. Zalizniak's category of adpositionality, contra Bulygina 1977) but still carries some traces of its morphophonological origins (partial lexical selectivity, adjacency requirement). Further development leads to definitive de-morphophonologization: the inclusion of comparative forms and availability of distant control. In the long run, this may lead to a situation where the trigger of a variant becomes a true controller of a morphosyntactic value, the initial stem alternation turning into a full-fledge inflectional category[6].

More on the dynamics of this evolution can only be learned through an analysis of additional sources—such as internet blogs—because the statistics of innovative (first of all distant) uses of 'n'-forms in the Russian National Corpus, on which this paper is based, are far too poor for any substantial conclusions.

References

1. *Lunt H.* (2001), Old Church Slavonic Grammar, Mouton de Gruyter, New York.
2. *Mattissen J.* (2003), Dependent-Head Synthesis in Nivkh, Benjamins, Amsterdam.
3. *Bulygina T. V.* (1977), Issues in the theory of morphological models [Problemy teorii morphologicheskikh modeley], Nauka, Moscow.
4. *Valova E.* (2012), The government of the prepositions «между», «меж», «промежду», «промеж» in Standard Russian of the XVIII through XXI centuries. [Upravlenie predlogov «между», «меж», «промежду», «промеж» v russkom literaturnom yazyke XVIII–XXI vekov], manuscript.
5. *Yemtseva K.* (2011), Pronominal forms in 'n': a study based on the Russian National Corpus [«Prikrytye» formy mestoimeniy: issledovanie na materiale Natsionalnogo korpusa russkogo yazyka], manuscript.
6. *Zalizniak A. A.* (1967), Russian nominal inflection [Russkoe imennoe slovoizmeneniye], Nauka, Moscow.
7. *Itkin I. B.* (2007), Russian morphophonology [Russkaja morfonologija], Gnozis, Moscow.
8. *Polivanova A. K.* (2014), Old Church Slavonic: grammar and vocabularies [Staroslavjanskij jazyk: grammatika, slovari], Universitet Dmitriya Pozharskogo, Moscow.
9. *Shakhmatov A. A.* (1957), Diachronic morphology of Russian [Istoricheskaya morfologiya russkogo yazyka], Uchpedgiz, Moscow.

[6] This would be not unlikely the evolution of e.g. initial vocative lenition in some Celtic languages (where the category status is however obtained by the loss of controller, whereby the lenition becomes a grammatical category per se).

МОДАЛЬНЫЕ ЧАСТИЦЫ И ИДЕЯ АКТУАЛИЗАЦИИ ЗАБЫТОГО (НА МАТЕРИАЛЕ ПАРАЛЛЕЛЬНЫХ КОРПУСОВ)[1]

Добровольский Д. О. (dm-dbrv@yandex.ru),
Левонтина И. Б. (irina.levontina@mail.ru)

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

В русском языке имеется богатый репертуар дискурсивных средств для выражения идеи актуализации забытого. В одних случаях говорящий напоминает адресату о каком-то объекте или событии, в других — говорящий пытается припомнить какую-то деталь или название, касающиеся событий, о которых он рассказывает. Некоторые дискурсивные слова обслуживают оба типа контекстов, другие специализируются на одном из них. Мы рассмотрели также переводные эквиваленты, которые используются для этих слов в английском и немецком языках. Мы опирались на данные параллельных корпусов НКРЯ. В английском соответствующие значения если и выражаются, то скорее либо синтаксически, либо с помощью эксплицитных высказываний. В немецком же репертуар дискурсивных средств не менее богат, чем в русском, однако при этом нет одно-однозначного соответствия между русскими и немецкими дисурсивными словами. Семантические конфигурации у них различные, и хотя смысловые компоненты зачастую очень похожи, но сочетается по-разному, и поэтому в разных контекстах все частицы переводятся достаточно разнообразно.

Ключевые слова: корпусная лингвистика, параллельные корпусы, контрастивные исследования, модальные частицы, семантика, перевод

MODAL PARTICLES AND THE ACTUALIZATION OF FORGOTTEN DETAILS (BASED ON THE MATERIALS OF PARALLEL CORPORA)

Dmitrij Dobrovol'skij (dm-dbrv@yandex.ru),
Irina Levontina (irina.levontina@mail.ru)

RLI RAS, Moscow, Russia

[1] Работа выполнена при поддержке РФФИ (грант 13-06-00403) и РГНФ (грант 15-04-12018).

The use of parallel corpora carries with it special problems, particularly when it comes to units that are typical of oral speech. Nevertheless, it is the presence of good Russian-English and Russian-German parallel texts in the RNC that has made the present study possible. Our analysis also demonstrates the limitations inherent in investigations based on parallel corpora, especially with respect to discursive words. Only a combination of various research methods is capable of producing adequate results.

In this study we analyze a group of discursive words whose semantics actualize things that have been forgotten. There are two types of situations here. In the first, the speaker reminds the addressee of some object or event; in the second, the speaker attempts to remember some detail or name connected with the events s/he is talking about. Certain discursive words apply to both types of contexts, whereas others are special to one of them. Among the Russian discursive words that express these ideas are units such as *biš', tam, ešče, pomnite, ėtot* and phrasemes like *kak tam, kak ego, kak že, ėto samoe*, etc. We also examine English and German equivalents used to translate these words.

Russian has a rich repertory of discursive resources for actualizing forgotten details. In English, if the corresponding meanings are expressed at all, it tends to be done either syntactically or by means of explicit utterances. The German arsenal of discursive resources is no less extensive than the Russian, but there are no one-to-one correspondences between the Russian and German discursive words. They have different semantic configurations, and although the meaning components are often quite similar, they combine differently, so that translations of such particles in various contexts are rather diverse.

Key words: corpus linguistics, parallel corpora, contrastive studies, modal particles, semantics, translation

1. Использование современных технологий, в частности электронных корпусов, с одной стороны, существенно облегчают работу лингвиста, предоставляя в его распоряжение практически неограниченный объем данных. Однако с другой стороны, действительно эффективное использование новых технологий невозможно без их теоретического осмысления.

Корпусный материал активно используется в современных лингвистических исследованиях. Возникает впечатление, что само по себе обращение к корпусам является гарантией объективности исследования и представительности эмпирических данных. В действительности же всегда остается вопрос, насколько полученные данные отражают реальность языка и соответствуют интуиции его носителей.

Для получения достоверного представления о функционировании языковых единиц необходимо сочетание разных методов, таких как: опросы информантов и инроспекция. В простых случаях при корректном использовании разных методов они дают сходные результаты. Если же результаты применения разных методик существенно различаются, необходимо установить источник этого несоответствия. Иными словами, когнитивная реальность корпусных данных нуждается в дополнительных обоснованиях и эмпирической проверке. Априорно нельзя утверждать, что результаты анализа корпуса (даже если он достаточно большой и претендует на репрезентативность) отражает интуицию

носителей языка. Иными словами, языковая интуиция ("patterns of knowledge") не может быть приравнена к "patterns of use" (как они, в частности, фиксируются в корпусах). Ср., в частности, [Wray 2002: 277]. Теоретически возможно, что каждый из этих феноменов в какой-то степени существует независимо от другого. Ср. указание на то, что "it is still true that the relation between the two types of data remains unclear and that identity cannot be taken for granted" [Gilquin, Gries 2009: 17]. Таким образом, может оказаться, что корреляции между корпусными данными и языковой интуицией носителя довольно сложные.

По корпусным методам и их соотношению с экспериментальными данными (данными опросов), интроспекцией и пр. есть уже большая литература; ср., например, [Bybee, Hopper 2001; Nordquist 2004; Baroni, Guevara, Pirrelli 2007; Marzo, Rube, Umbreit 2007; Wulff 2007; McGee 2009]. Этой проблематике посвящен специальный номер журнала Corpus Linguistics and Linguistic Theory 5 (1) 2009 "Corpora and experimental methods". Ср. также [Gries, Hampe, Schönefeld 2005; Kepser, Reis 2005; Arppe, Järvikivi 2007; Mollin 2009; Divjak, Gries 2012]. Так, в работе [Littlemore, MacArthur 2012] разбирается показательный пример со словом thread: оказывается, что соотношение частоты разных значений этого слова по данным корпусов и по данным опросов радикально расходится. О соотношении разных методов применительно к исследованию русских дискурсивных слов и их переводных эквивалентов см. [Добровольский, Левонтина 2009)].

Использование параллельных корпусов сопряжено с дополнительными трудностями. Во-первых, их объем пока невелик. Так, входящий в состав НКРЯ русско-английский корпус включает около 3,6 миллионов словоупотреблений, русско-немецкий — 2,7 миллионов, а другие русско-иноязычные корпусы существенно меньше. Во-вторых, они несбаллансированы: в них входит в основном художественная литература (проза), почти исключительно литература XIX века (причем очень ограниченный набор авторов).

Понятно, что для наших целей — описания особенностей употребления дискурсивных слов — этот материал совершенно недостаточен. Впрочем, параллельные корпусы устной речи — это задача, которая в принципе непонятно, как могла бы быть решена. Для описания дискурсивных слов было бы полезно иметь хотя бы тексты пьес и тех прозаических произведений, в которых много диалогов, причем это должны быть в первую очередь более современные произведения. При этом ясно, что реальная спонтанная устная речь и ее имитация в художественной литературе — это разные вещи.

В современных корпусных исследования большие возможности сулит также использование так называемых «сравнимых (comparable)» корпусов, на что нам указал один из анонимных рецензентов. Однако как раз для исследования дискурсивных слов корпусы этого типа едва ли могу предоставить какой-либо материал.

Несмотря на все это, анализ, опирающийся не только на инроспекцию, но и на данные параллельных корпусов, может дать интересные результаты.

В данной работе мы хотим проанализировать группу дискурсивных слов с семантикой актуализации забытого. Речь идет о двух типах ситуаций. В одних случаях говорящий напоминает адресату о каком-то объекте или событии,

в других — говорящий пытается припомнить какую-то деталь или название, касающиеся событий, о которых он рассказывает. Некоторые дискурсивные слова обслуживают оба типа контекстов, другие специализируются на одном из них.

К русским дискурсивным словам, выражающим эти идеи, относятся такие единицы, как *бишь* (*Где бишь мой рассказ несвязный?*), *там* (*Какой там у него номер дома?*), *еще* (*Ты с ней спал еще*), *помните* (*Вы про нее, помните, мне рассказывали?*), *этот* (*Эту… как ее… переписку Энгельса с этим… как его, дьявола… с Каутским*[2]), а также обороты *как там* (*Это было в журнале, как там, «Вопросы языкознания»*), *как его, как же, это самое* и др. Мы рассмотрим также переводные эквиваленты, которые используются для этих слов в английском и немецком языках. Мы будем опираться на данные параллельных корпусов НКРЯ в той мере, в какой это окажется возможным.

2. Наиболее ярким словом этой семантики является *бишь*[3].

В современном языке *бишь* — вопросительная частица, которая используется (хотя и постепенно уходит из активного употребления) в сочетании с вопросительным местоимением, в постпозиции к нему, для указания на то, что запрашиваемая информация была доступна говорящему еще недавно, но вдруг выпала из его сознания, а она нужна ему для обеспечения связности текста. Невозможны, во всяком случае в современном языке, сочетания типа **Сколько бишь будет семью восемь?; ??В каком бишь году была Куликовская битва?* Хотя эти сведения вполне могут быть забыты, *бишь* здесь неуместно: оно обычно связано с чем-то, забытым непосредственно в процессе разговора, или с каким-то сведением — а чаще всего наименованием — которое понадобилось говорящему в ходе этого разговора. *Бишь* не вполне уместно, если человек пытается извлечь из глубин памяти что-то давно и прочно забытое. В этом случае он скорее скажет, например, не *Как* ***бишь*** *ее звали?*, а *Как* ***же*** *ее звали*?

Очень показательно, что во многих случаях частица *бишь* сочетается с такими показателями внезапного вспоминания, как *А!, Да!* и такими средствами восстановления связности текста, как *так, так вот* и др. См. подробнее [Левонтина 2011; 2014].

На английском *бишь* в подобных контекстах чаще всего остается без перевода:

(1) *Не отрицаю, впрочем, что мне теперь гораздо лучше. Да, так на чем,* ***бишь****, я остановился? Мороз, эти летящие трамваи*. [М. А. Булгаков. Мастер и Маргарита (1929–1940)]
I don't deny, however, that I'm much better now. Yes, so where did I leave off? Frost, those flying trams… [Mikhail Bulgakov. Master and Margarita (Richard Pevear, Larissa Volokhonsky, 1979)]

---

[2] М. А. Булгаков. Собачье сердце.

[3] *Бишь* также употребляется в составе союза *то бишь*, который имеет совсем другое значение (= *то есть*) и который мы здесь не рассматриваем.

(2) *— Да, о чем* ***бишь*** *я думал? — спросил себя Нехлюдов, когда все эти перемены в природе кончились.* [Л. Н. Толстой. Воскресение (1899)]
"Why, what was I thinking about?" Nekhludoff asked himself when all these changes in nature were over. [Leo Tolstoy. The Awakening (William E. Smith, 1900)]

В английской устной речи сходную функцию может выполнять использование прошедшего времени вместо настоящего: *What was your name?* Кстати, такое использование претерита характерно и для немецкого языка.[4]

В немецком есть и ряд лексических средств выражения этого значения. Это в первую очередь такие частицы, как *doch, noch, gleich, halt, denn, eben*, *nur* или наречие *noch einmal* (*Wie war das* ***noch einmal****?; Wie war* ***noch einmal*** *der Name?*). Причем они могут использоваться в самых разных сочетаниях, ср. *doch noch, doch gleich, denn noch, halt doch eben*. Однако некоторые сочетания оказываются невозможными. Например, *gleich* и *noch,* выражающие одну и ту же идею, никогда не употребляются вместе.

Приведем некоторые примеры:

(3) *Коля говорил, что вас князь ребенком назвал… это хорошо… Да, что* ***бишь*** *я… еще что-то хотел…* [Ф. М. Достоевский. Идиот (1868–1869)]
Kolja hat mir gesagt, der Fürst habe Sie ein Kind genannt… Das ist eine richtige Bezeichnung… Ja, was hatte ich ***doch noch****… ich wollte noch etwas sagen…* [Fëdor Michajlovič Dostojewski. Der Idiot (Hermann Röhl, 1981)]

(4) *Третье, — что* ***бишь*** *еще ты сказал? Князь Андрей загнул третий палец.* [Л. Н. Толстой. Война и мир, том 2 (1865–1869)]
Drittens ja, was sagtest du ***denn noch****? — Fürst Andrej bog den dritten Finger um.* [Leo Tolstoj. Krieg und Frieden, 2. Band (Hermann Röhl, 1922)]

(5) *Хоть назад бы доехали в Гришкино, ночевали бы у Тараса. А то вот сиди ночь целую. Да что,* ***бишь****, хорошего было?* [Л. Н. Толстой. Хозяин и работник]
Und selbst wenn wir wieder nach Grischkino zurückgekommen wären, so hätten wir wenigstens bei Taras übernachten können. Aber nun kann man hier die ganze Nacht sitzen. Ja, ich dachte doch vorhin an etwas Schönes; was war das ***nur****?* [Leo Tolstoj. Herr und Knecht]

Наиболее интересны случаи вынесения выражения, эквивалентного *бишь*, в отдельное высказывание (в английском в таких случаях используется клефт):

---

[4] Для русского такое использование прошедшего времени менее характерно, хотя в какой-то степени прошедшее время может передавать сходную идею. Ср. «*У него*, по-моему, *был брат, — сказал Сталин, — интересно, где он сейчас?*» [Ф. А. Искандер. Сандро из Чегема].

(6) *«Да, что **бишь** еще неприятное он пишет? — вспоминал князь Андрей содержание отцовского письма.* [Л. Н. Толстой. Война и мир, том 2 (1865–1869)]
***Ja, was war es doch**? Schrieb er nicht auch etwas Unangenehmes? dachte Fürst Andrej und rief sich den Inhalt des Briefes noch einmal ins Gedächtnis zurück.* [Leo Tolstoj. Krieg und Frieden, 2. Band (Hermann Röhl, 1922)]

Как мы видим, разные переводные эквиваленты как бы выхватывают разные части толкования слова *бишь*. *Halt* выражает идею ‘остановись, мне нужно время’ (в случае *бишь* говорящий забыл и пытается вспомнить, из-за чего происходит запинка в коммуникации); *gleich* и сходное с ним в этом значении *eben* (‘только что’) вводит идею ‘только что я это знал, помнил’; *noch* (‘еще’) и соответствует идее восстановления связности текста. Наконец, о частице *doch* можно сказать, что она как бы дает коммуникативный пинок. Сама по себе она идею *бишь* не передает, однако очень уместна в сочетании с другими.

Что касается английского языка, то, как уже было сказано, лексическое выражение рассматриваемой функции для него нехарактерно. Однако она успешно выполняется синтаксическими средствами — с помощью так называемого клефта (расщепленного предложения):

(7) *О чем **бишь** я спорил? — думал он.* [Л. Н. Толстой. Анна Каренина (1878)]
“***Whatever was it** I was disputing about?” he wondered.* [Leo Tolstoy. Anna Karenina (Constance Garnett, 1911)].

Отдельно можно выделить форму *как бишь его* в функции припоминания имени человека. В этом случае стандартным английским переводом оказывается *what's his name*:

(8) *Ей все этот старичок рассказал… **как бишь его**?* [М. Ю. Лермонтов. Герой нашего времени (1839–1841)]
*That little old man — **what's his name** — has told her everything.* [Mikhail Lermontov. A Hero of Our Time (J. H. Wisdom, Marr Murray, 1916)]

Само по себе выражение *what's his name?* на первый взгляд не содержит никаких дискурсивных маркеров. Интересно, однако, его превращение в устойчивый оборот, который произносится как единая синтагма и может субстантивироваться. Английские словари даже выделяют форму *what's-his-name* в качестве особого местоимения со значением близким к русскому *этот как его*:

(9) *А этот, как его, он турок или грек? Тот черномазенький, на ножках журавлиных, / Не знаю, как его зовут, / Куда ни сунься: тут как тут* [А. С. Грибоедов. Горе от ума].

Интересно, что имеется некоторое различие между русским *как-там-тебя* и английским *what's-your-name*. Последнее может использоваться не только

тогда, когда говорящий забыл некое имя, но и тогда, когда оно никогда не было ему известно.[5] Так, если говорящий стоит на берегу и видит незнакомого человека на борту своей яхты, он крикнет *Hey, girl, what's-your-name?* (в меньшей степени это относится и к форме *what's-his-name*). На это обстоятельство обратила наше внимание анонимный рецензент Т. Е. Янко. Она также отметила, что «по сообщению носителей английского языка это выражение произносится с акцентом на *what*, что, вообще говоря, не соответствует законам интонирования предложения с такой же структурой *What is your name?*».

3. Теперь остановимся кратко на другой частице, которая в одном из своих многочисленных значений сближается с рассмотренной частицей *бишь*. Речь идет о частице *там*; ср. *Какой* ***там*** *у него адрес?; Как* ***там*** *его по отчеству?* В основе этого значения там лежит яркая внутренняя форма этого слова, его пространственное значение. Говорящий как бы удаляет от себя некоторый объект, либо дистанцируясь от него, либо убирая его из фокуса внимания: *Какой* ***там*** *у него телефон?; Как* ***там*** *у Пушкина?* Говорящий представляет информацию не столько как неизвестную (как в случае *Какой у него телефон?*), сколько как находящуюся далеко, за пределами светлого поля сознания и потому не вполне доступную. Отметим, что в нашем случае *там* просодически оформляется как клитика, в отличие от некоторых других *там*, имеющих просодию вводного слова. Использование идеи отдаления для указания на то, что что-то было забыто, очень естественно. Ср. выше об употреблении прошедшего времени вместо настоящего в английском и немецком (отдаление во времени).

Можно отметить, что *там* имеет ряд тонких отличий от *бишь*.

Там гораздо менее привязано к контексту конкретного разговора; оно, в отличие от *бишь*, совершенно свободно используется во фразах типа *Сколько* ***там*** *будет семью восемь?; В каком* ***там*** *году была Куликовская битва?*

По этой же причине невозможны фразы типа **О чем* ***там*** *я?*, чрезвычайно характерные для *бишь*. В контексте восстановления нарушенной связности идея отдаления неуместна. Здесь как раз хорошо подходит другое дейктическое слово — *это*: *О чем* ***это*** *я?* Возможно, однако *Что* ***там*** *ты говорил насчет…?; О чем* ***там*** *мы говорили?* — здесь происходит возвращение к уже закрытой теме. Связность как раз не была нарушена, маркируется смена темы.[6]

Это *там* также может оставаться без перевода не только в английском, но и в немецком. Или могут использоваться отдаленные соответствия, такие как *oder so ähnlich* (≈ *или что-то в этом роде*):

(10) *По поводу всех этих вопросов, преступлений, среды, девочек мне вспомнилась теперь, — а впрочем, и всегда интересовала меня, — одна ваша статейка: «О преступлении» …* ***или как там*** *у вас, забыл название,*

---

[5] См. об этом выражении [Zwicky 1974].

[6] См. подробнее в [Левонтина 2011].

не помню. [Ф. М. Достоевский. Преступление и наказание (1866)]
Anläßlich aller dieser Fragen, Verbrechen, kleiner Mädchen, des Milieus ist mir eben ein kleiner Aufsatz von Ihnen eingefallen — er hat mich übrigens immer interessiert. Er heißt ›Vom Verbrechen‹ ***oder so ähnlich****, ich weiß nicht mehr genau*. [Fjodor Dostojewski. Verbrechen und Strafe (Alexander Eliasberg, 1924)]

В принципе, для перевода *там* на немецкий могут использоваться те же частицы, что и для перевода *бишь*. Существенно, однако, что при этом *gleich* и *eben* будут здесь наименее удачны, потому что в них не только отсутствует идея отдаления, лежащая в основе *там*, но и наоборот, фокусируется идея близости (временной), характерная именно для *бишь*.

4. Надо заметить, что русские *помните* и *помнишь* имеют вводное употребление, в котором сближаются с рассмотренными выше дискурсивными частицами. При этом в русском языке два варианта актуализации забытого — ориентированные на говорящего и на адресата (припоминание и напоминание) — достаточно четко противопоставлены, однако в переводах они часто отражаются одинаково.

(11) *Ну можно ли было предполагать, когда,* ***помните****, Чичиков только что приехал к нам в город, что он произведет такой странный марш в свете?* [Н. В. Гоголь. Мертвые души (1835–1852)].[7]

Помните часто переводится буквально, но интересно, что в немецких примерах еще добавляются частицы *doch noch*.

(12) *— Это вот та старуха, — <…> — та самая, про которую,* ***помните****, когда стали в конторе рассказывать, а я в обморок-то упал.*
[Ф. М. Достоевский. Преступление и наказание (1866)]
»Es ist dieselbe Alte,« <…> »Dieselbe, von der man, ***Sie erinnern sich doch noch****, im Polizeibureau zu sprechen anfing, worauf ich in Ohnmacht fiel.«* [Fjodor Dostojewski. Verbrechen und Strafe (Alexander Eliasberg, 1924)]

(13) *Так проходя-то в восьмом часу-с, по лестнице-то, не видали ль хоть вы, во втором-то этаже, в квартире-то отворенной —* ***помните****?*
[Ф. М. Достоевский. Преступление и наказание (1866)]
Nun, als Sie gegen acht über die Treppe gingen, sahen Sie da nicht im ersten Stock, in einer offen stehenden Wohnung — ***Sie erinnern sich doch noch****?* [Fjodor Dostojewski. Verbrechen und Strafe (Alexander Eliasberg, 1924)].

Или частицы *schon* и *noch* (ср. примеры 14 и 16).

---

[7] Кстати, английский переводчик (D. J. Hogarth, 1931) вообще выпустил тут целый кусок текста, оставив только первую фразу фрагмента.

(14) *Ну и прекрасно; радоваться всегда не худо. А к той,* ***помнишь****? послал?* [И. С. Тургенев. Отцы и дети (1860–1861)]
Herrlich! Es ist immer gut, sich zu freuen. Aber hat man dort hingeschickt? ***Du weißt schon.*** [Iwan Turgenew. Väter und Söhne (1911)]

Сходным образом это происходит в английских переводах:

(15) *Как же! мы виделись у Росси,* ***помните****, на этом вечере, где декламировала эта итальянская барышня — новая Рашель.* [Л. Н. Толстой. Анна Каренина (1878)]
To be sure! We met at Rossi's, ***do you remember****, at that soiree when that Italian lady recited — the new Rachel?* [Leo Tolstoy. Anna Karenina (Constance Garnett, 1911)]

Ср. также следующие примеры из русских текстов с английским и немецким переводами:

(16) *Тут один стервец у меня сидит, мальчишка. Понимаешь, на станции попался тот самый Жухрай,* ***помнишь****, который железнодорожников натравил на нас.* [Н. А. Островский. Как закалялась сталь (ч. 1) (1930–1934)]
I've got a damn nuisance of a boy here. ***Remember*** *that chap Zhukhrai, the one who stirred up the railway-men against us? Well, he was caught at the station.* [Nikolai Ostrovsky. How the Steel was Tempered (part 1) (R. Prokofieva, 1952)]
Ich habe da so einen niederträchtigen Halunken zu fassen gekriegt, einen Rotzkerl. Verstehst du, uns war eben der Shuchrai in die Hände gefallen, der — ***weißt du noch?*** *—, der die Eisenbahner gegen uns aufgehetztha*t. [Nikolai Ostrowski. Wie der Stahl gehärtet wurde (erster Teil) (1936–1977)]

(17) *А ведь признайся, брат, ведь ты, право, преподло поступил тогда со мною,* ***помнишь****, как играли в шашки.* [Н. В. Гоголь. Мертвые души (1835–1852)]
However, my friend, you must admit that you treated me rather badly the day that we played that game of chess… [Nikolay Gogol. Dead Souls (D. J. Hogarth, 1931)]
Aber gib's nur zu, Bruder, das war damals wirklich gemein von dir, ***erinnerst du dich****, als wir Dame gespielt haben, ich hatte doch gewonnen… Ja, Bruder, du hast mich einfach übers Ohr gehauen*. [Nicolaj Gogol. Die toten Seelen (Michael Pfeiffer, 1978)]

При этом дискурсивное *remember* нередко сопровождается модальными словами (*may, will*):

(18) *Я ему сулил каурую кобылу, которую,* ***помнишь****, выменял у Хвостырева*. [Н. В. Гоголь. Мертвые души (1835–1852)]
I have promised him the roan filly which, as ***you may remember****, I swopped from Khvostirev».* [Nikolay Gogol. Dead Souls (D. J. Hogarth, 1931)]

(19) *Я ведь тебе говорил,* ***помнишь****, про одну барышню, в которую я был влюблен, бывши ребенком.* [Л. Н. Толстой. Юность (1856)]
You will remember *how I told you about a girl with whom I used to be in love when was a little boy?* [Leo Tolstoy. Youth (C. J. Hogarth, 1910–1935)]

5. В невопросительных контекстах в русском языке также используется интересный маркер напоминания — частица *еще*; ср. *у нее* ***еще*** *юбка была клетчатая*. Такое *еще* не используется просто при рассказе о человеке: *Я увидел девушку с большой синей сумкой и в темных очках. *У нее* ***еще*** *юбка была клетчатая.* Смысл этого *еще* в том, чтобы уточнить неудавшуюся идентификацию объекта. Говорящий уже назвал какие-то отличительные особенности, но видит, что их недостаточно. Здесь проявляется связь с основным значением *еще* — говорящий **дополняет** уже сделанное описание. Интересно, однако, что *еще* обычно используется именно в контексте напоминания, а не, например, при идентификации одного из множества видимых объектов: *?Вон, смотри, она в левом верхнем углу, у нее* ***еще*** *юбка клетчатая.* Поэтому для *еще* характерны сочетания с *помнишь, помните, вспомни:*

(20) *Помнишь нашу соседку? Она* ***еще*** *в обувном работает. Я поразился, когда на Новый год увидел в школе её дочь.* [«Знание — сила», 2003]

(21) *Помнишь, Постум, у наместника сестрица? / Худощавая, но с полными ногами. / Ты с ней спал* ***еще****… Недавно стала жрица. / Жрица, Постум, и общается с богами.* [И. Бродский]

Это как раз тот случай, когда параллельные корпусы в их нынешнем состоянии мало помогают в установлении переводных эквивалентов: мы имеем дело с не самым частым и очень разговорным значением частого слова. Даже само по себе увеличение корпуса здесь не поможет: нужна тонкая семантическая разметка, чтобы отделить нужное значение таких слов, как *еще, noch* и под. от всех других.

Можно, однако, отметить, что для немецкого также характерно использование в аналогичных контекстах частицы *noch* (*еще*):

(22) *…und du hast noch mit ihm eine geraucht*
[http://www.youtube.com/watch?v=6uKvy4qs61I]

Этот пример интересен порядком слов. Дело в том, что, если бы имелось в виду, что адресат выкурил еще одну сигарету, то *noch* стояло бы непосредственно перед *eine: Und du hast mit ihm noch eine geraucht*. При имеющемся же порядке слов возможно только одно понимание: ‘ты еще с ним однажды покурил’. Таким образом, в этом значении *noch* действительно очень похоже на рассматриваемое значение *еще*.

6. Сама природа идеи актуализации забытого такова, что очень естественным оказывается использовать соответствующие слова как заполнители пауз хезитации. Особенно это характерно для сочетаний *как там, как там его, как его*.

Встретившееся в цитированном выше примере из Достоевского (*как там у вас*) *как там* в русском языке в значительной степени лексикализовалось и может иметь вводное употребление (*Он учится на факультете, как там, защиты информации*).

Когда речь идет об имени человека, в русской устной речи часто используется выражение *как его*. В этом случае досточно стандартный перевод на немецкий — *wie heißt er doch gleich,* а на английский — *what's his name*:

(23) — *У вас ведь, кажется, только еще одна комната и занята.* ***Этот, как его*** *Ферд… Фер… — Фердыщенко.* [Ф. М. Достоевский. Идиот (1868–1869)]
»Es ist bei euch, soviel ich weiß, erst ein Zimmer vermietet. An diesen, ***wie heißt er doch gleich****? Ferd… Fer… « » Ferdyschtschenko.»* [Fëdor Michajlovič Dostojewski. Der Idiot (Hermann Röhl, 1981)]

(24) *Да не только мышь, не проникнет даже* ***этот, как его****…из города Кириафа.* [М. А. Булгаков. Мастер и Маргарита (1929–1940)]
Not only a mouse, but even that one, ***what's his name****…from the town of Kiriath, couldn't get through.* [Mikhail Bulgakov. Master and Margarita (Richard Pevear, Larissa Volokhonsky, 1979)]

Итак, мы увидели, что в русском языке имеется богатый репертуар дискурсивных средств для выражения идеи актуализации забытого. В английском соответствующие значения если и выражаются, то скорее либо синтаксически, либо с помощью эксплицитных высказываний. В немецком же репертуар дискурсивных средств не менее богат, чем в русском, однако при этом нет одно-однозначного соответствия между русскими и немецкими дисурсивными словами. Семантические конфигурации у них различные, и хотя смысловые компоненты зачастую очень похожи, но сочетаются по-разному, и поэтому в разных контекстах все частицы переводятся достаточно разнообразно.

Наша работа стала возможной благодаря наличию в рамках НКРЯ хороших корпусов русско-английских и русско-немецких параллельных текстов. Предсказать все богатство возможных межъязыковых соответствий невозможно. Вместе с тем, наш анализ продемонстрировал и ограничения, свойственные исследованиям, основанным на параллельных корпусах, особенно применительно к дискурсивным словам. Только комбинация разных методов исследования может дать адекватный результат.

Литература

1. *Добровольский Д. О., Левонтина И. Б.* (2009), 500 способов сказать «нет» (русско-немецкие соответствия) // Логический анализ языка. Ассерция и негация / Отв. ред. член-корр. РАН Н. Д. Арутюнова. М.: Индрик, 2009. С. 400–410.

2. *Левонтина И. Б.* (2011), Частицы переспроса и припоминания // Слово и язык. Сборник статей к восьмидесятилетию академика Ю. Д. Апресяна. М. Языки славянских культур, 2011. С. 269–278.
3. *Левонтина И. Б.* (2014), Дискурсивные слова в вопросительных предложениях // Die Welt der Slaven, Vol. LX, pp. 201–218.
4. *Arppe A., Järvikivi J.* (2007), Every method counts: Combining corpus-based and experimental evidence in the study of synonymy, Corpus Linguistics and Linguistic Theory 3 (2), pp. 131–159.
5. *Baroni M., Guevara E., Pirrelli V.* (2007), Generating well-formed compounds: A corpus-based model tested against psycholinguistic evidence, available at: http://ucrel.lancs.ac.uk/publications/CL2007/paper/205_Paper.pdf
6. *Bybee J., Hopper P.* (2001), Introduction to frequency and the emergence of linguistic structure, Frequency and the emergence of linguistic structure, 1–24, John Benjamins, Amsterdam.
7. *Divjak D., Gries S. T.* (eds.). (2012), Frequency effects in language representation, Mouton de Gruyter, Berlin, New York.
8. *Gilquin G., Gries S. T.* (2009), Corpora and experimental methods: A state-of-the-art review, Corpus Linguistics and Linguistic Theory 5 (1), pp. 1–26.
9. *Gries S. T., Hampe B., Schönefeld D.* (2005), Converging evidence: Bringing together experimental and corpus data on the association of verbs and constructions, Cognitive Linguistics 16 (4), pp. 635–676.
10. *Kepser S., Reis M.* (eds.) (2005), Linguistic evidence: Empirical, theoretical and computational perspectives, Mouton de Gruyter, Berlin, New York.
11. *Littlemore J., MacArthur F.* (2012), Figurative extensions of word meaning: How do corpus data and intuition match up?, Frequency effects in language representation, Mouton de Gruyter, Berlin, New York, Vol. 2, pp. 195–233.
12. *Marzo D., Rube V., Umbreit B.* (2007), Salience and frequency of meanings: Comparison of corpus and experimental data on polysemy, available at: http://ucrel.lancs.ac.uk/publications/CL2007/paper/205_Paper.pdf
13. *McGee I.* (2009), Adjective-noun collocations in elicited and corpus data: Similarities, differences, and the whys and wherefores, Corpus Linguistics and Linguistic Theory 5 (1), pp. 79–103.
14. *Mollin S.* (2009), Combining corpus linguistic and psychological data on word co-occurrences: Corpus collocates versus word associations, Corpus Linguistics and Linguistic Theory 5 (2), pp. 175–200.
15. *Nordquist D.* (2004), Comparing elicited data and corpora, Language, culture and mind, CSLI Publications, Stanford, CA, pp. 211–223.
16. *Wray A.* (2002), Formulaic language and the lexicon, Cambridge University Press, Cambridge.
17. *Wulff S.* (2007), Combining corpus and experimental data to capture idiomaticity, available at: http://ucrel.lancs.ac.uk/publications/CL2007/paper/205_Paper.pdf
18. *Zwicky A.* (1974), Hey, Whatsyourname!, Papers from the Tenth Regional Meeting of the Chicago Linguistics Society, Chicago, pp 787–801.

References

1. *Arppe A., Järvikivi J.* (2007), Every method counts: Combining corpus-based and experimental evidence in the study of synonymy, Corpus Linguistics and Linguistic Theory 3 (2), pp. 131–159.
2. *Baroni M., Guevara E., Pirrelli V.* (2007), Generating well-formed compounds: A corpus-based model tested against psycholinguistic evidence, available at: http://ucrel.lancs.ac.uk/publications/CL2007/paper/205_Paper.pdf
3. *Bybee J., Hopper P.* (2001), Introduction to frequency and the emergence of linguistic structure, Frequency and the emergence of linguistic structure, 1–24, John Benjamins, Amsterdam.
4. *Divjak D., Gries S. T.* (eds.). (2012), Frequency effects in language representation, Mouton de Gruyter, Berlin, New York.
5. *Dobrovolskij D. O., Levontina I. B.* (2009), 500 ways to say "no" (Russian-German correlations) [500 sposobov skazat' "net" (russko-nemetskie sootvetstviya)], Logical analysis of language. Assertion and negation [Logicheskiy analiz yazyka. Assertsiya i negatsiya], Indrik, Moscow, pp. 400–410.
6. *Gilquin G., Gries S. T.* (2009), Corpora and experimental methods: A state-of-the-art review, Corpus Linguistics and Linguistic Theory 5 (1), pp. 1–26.
7. *Gries S. T., Hampe B., Schönefeld D.* (2005), Converging evidence: Bringing together experimental and corpus data on the association of verbs and constructions, Cognitive Linguistics 16 (4), pp. 635–676.
8. *Kepser S., Reis M.* (eds.) (2005), Linguistic evidence: Empirical, theoretical and computational perspectives, Mouton de Gruyter, Berlin, New York.
9. *Levontina I. B.* (2011), Particles of overinterrogation and remembering [Chastitsy peresprosa i pripominaniya], Word and language [Slovo i yazyk], Yazyki slavianskikh kul'tur, Moscow, pp. 269–278.
10. *Levontina I. B.* (2014), Discursive words in interrogative sentences [Diskursivnye slova v voprositel'nykh predlozheniyakh], Die Welt der Slaven, Vol. LX, pp. 201–218.
11. *Littlemore J., MacArthur F.* (2012), Figurative extensions of word meaning: How do corpus data and intuition match up?, Frequency effects in language representation, Mouton de Gruyter, Berlin, New York, Vol. 2, pp. 195–233.
12. *Marzo D., Rube V., Umbreit B.* (2007), Salience and frequency of meanings: Comparison of corpus and experimental data on polysemy, available at: http://ucrel.lancs.ac.uk/publications/CL2007/paper/205_Paper.pdf
13. *McGee I.* (2009), Adjective-noun collocations in elicited and corpus data: Similarities, differences, and the whys and wherefores, Corpus Linguistics and Linguistic Theory 5 (1), pp. 79–103.
14. *Mollin S.* (2009), Combining corpus linguistic and psychological data on word co-occurrences: Corpus collocates versus word associations, Corpus Linguistics and Linguistic Theory 5 (2), pp. 175–200.
15. *Nordquist D.* (2004), Comparing elicited data and corpora, Language, culture and mind, CSLI Publications, Stanford, CA, pp. 211–223.

16. *Wray A.* (2002), Formulaic language and the lexicon, Cambridge University Press, Cambridge.
17. *Wulff S.* (2007), Combining corpus and experimental data to capture idiomaticity, available at
http://ucrel.lancs.ac.uk/publications/CL2007/paper/205_Paper.pdf
18. *Zwicky A.* (1974), Hey, Whatsyourname!, Papers from the Tenth Regional Meeting of the Chicago Linguistics Society, Chicago, pp 787–801.

ПОКАЗАТЕЛЬ СОСЛАГАТЕЛЬНОГО НАКЛОНЕНИЯ КАК ЧАСТЬ СОЮЗА[1]

Добрушина Н. Р. (nina.dobrushina@gmail.com)

Национальный исследовательский университет
Высшая школа экономики, Москва, Россия

Статья посвящена союзам, содержащим частицу *бы (б)*. Рассматривается вопрос о том, является ли элемент *бы (б)* показателем сослагательного наклонения, и если да, то насколько автономным является этот показатель по отношению к союзу. Для решения первого вопроса применяются критерии дистрибуции форм в сфере действия союза и возможности повторения частицы. В статье делается вывод о том, что не содержат сослагательного наклонения сравнительные союзы *будто бы, как будто бы* и *словно бы*. Вопрос об автономности частицы *бы* решается на основании следующих критериев: может ли частица быть опущена и может ли она находиться на дистанции от союза. По этим признакам союзы разделились на несколько групп.

Ключевые слова: глагол, сослагательное наклонение, ирреалис, союз, энклитика

SUBJUNCTIVE PARTICLE AS A PART OF CONJUNCTION

Dobrushina N. R. (nina.dobrushina@gmail.com)

National Research University Higher School of Economics,
Moscow, Russia

Russian subjunctive is expressed by an analytical form which consists of subjunctive particle *by (b)* and past indicative or infinitive or a few predicative adverbs and adjectives. The subjunctive particle is an enclitic. It often merges with subordinate conjunctions, which yields words functioning as conjunctions and containing the subjunctive particle. Historically, the particle *by* in conjunctions can be traced back to the marker of subjunctive. Synchronically, however, the group is not homogenous. The aim of the paper

[1] Исследование выполнено в рамках проекта описания русской корпусной грамматики «Русграм» (http://rusgram.ru) при поддержке фонда РГНФ (Russian Foundation for Humanities), проект N 14-04-00264.

is to find out which of the conjunctions with *by* should be considered as containing the marker of subjunctive, and test whether the particle can or can not be separated from the conjunction. Four criteria are used. The first and the second, namely, (a) the forms available in the subordinate clause with the conjunction and (b) the possibility of repetition of the particle *by* with the second predicate shows that comparative conjunctions do not synchronically contain the subjunctive marker. The third and fourth criteria, namely (c) the omission of the particle *by* and (d) its ability to be separated from the conjunction by another words give different results.

Key words: verb, subjunctive, irrealis, conjunction, enclitic

1. Введение

В грамматических описаниях русского языка в списках союзов указываются *чтобы, как бы*, *будто бы*, *если бы, когда бы, лишь бы* и другие (РГ 1980: 473, 562; Апресян, Пекелис 2012). В тех же описаниях эти лексические единицы рассматриваются как сочетание союза и частицы *бы (б),* выражающей сослагательное наклонение, например: «...Союзы, включающие в свой состав частицу *бы* (*если бы*, прост. *кабы*, *добро бы*, устар. *диви бы*, *как если бы*, *чтобы*, устар. *дабы*), оказываются неотъемлемыми компонентами сложного комплекса „союз — глагольная форма на *-л*“, оформляющего синтаксические ирреальные наклонения» (Русская грамматика 1980: 541).

Итак, Русская грамматика называет частицу *бы* «компонентом» союза. Напротив, в статье Ричарда Брехта “Čtoby or čto and by” (Brecht 1977) высказано мнение, что в современном русском языке слово *чтобы* состоит из двух совершенно отдельных единиц *что* и *бы*: “...The historically independent *by* and *čto* have not undergone the process of univerbation in the modern language. In synchronic terms this means that each occurs completely independently of the other and their regular juxtaposition in pronunciation and orthography is the result of the fact that *by* acts like a clitic in the language” (Brecht 1977: 38).

В настоящей статье будут обсуждены следующие вопросы.

- Какие критерии могут быть применены для того, чтобы квалифицировать частицу *бы* в составе различных союзов (точнее было бы называть их союзными сочетаниями) как показатель сослагательного наклонения?
- В случае, если эти критерии позволяют принять решение о том, что частица является показателем сослагательного наклонения, можно ли считать союзное сочетание (*чтобы, как бы* и др.) полностью аналогичным двум отдельным словам (*что* и *бы, как* и *бы*)?

С этих точек зрения будут рассмотрены следующие единицы:
Чтобы, дабы, если бы, когда бы, кабы, как бы, *будто бы*, *как будто бы*, *словно бы, как если бы*, *добро бы, ладно бы, лишь бы, только бы, хотя бы, хоть бы, пускай бы*

Многие из этих единиц имеют несколько функций. Сочетание *как бы* может быть частицей (*Она* ***как бы*** *моя жена*), комбинацией союзного слова *как*

и частицы *бы* (*Думаю, **как бы** признаться*) и союзом (*Боюсь, **как бы** не узнали*). В этой работе будут рассматриваться только союзные употребления этих слов. Союзные употребления тоже могут быть разнообразны: союз *будто бы* используется для присоединения сетенциальных актантов (*Говорили, **будто бы** она его жена*) и сравнительного придаточного предложения (*Она ведет себя так, **будто бы** они уже женаты*). Объемы настоящей статьи не позволяют каждый раз обсуждать все употребления союза.

2. Критерии выделения сослагательного наклонения

Нам известны два формальных критерия, позволяющих установить, что частица *бы* является показателем сослагательного наклонения: дистрибуция форм в сфере действия частицы и возможность повторения частицы.

2.1. Дистрибуция форм

В независимых предложениях для частицы *бы* с ирреальным значением характерно сочетание с прошедшим временем, инфинитивом и некоторыми предикативами (*помог бы / помочь бы / надо бы*). В настоящей работе эти три сочетания во всех их контекстах считаются представителями сослагательного наклонения (эта позиция разделена не всеми лингвистами). Можно утверждать, что если в сфере действия лексемы, содержащей *бы*, регулярным образом возможны другие формы, то *бы* не является показателем сослагательного наклонения.

По этому критерию из всех единиц, которые были перечислены выше, лишь три союза не содержат частицу сослагательного наклонения, поскольку могут сочетаться с настоящим временем: это сравнительные союзы *будто бы, как будто бы* и *словно бы*.

(1) *И сложно поверить, **будто бы** она не **знает**, где спрятан артефакт.*
[Вячеслав Солдатенко (Слава Сэ). Ева (2010)][2]

(2) *К старости она так и осталась буквой «г», **как будто бы** она до сих пор **тянет** весь груз своей жизни на ручной тележке.* [«Бельские Просторы», 2010]

(3) *Волнуясь, он отчасти любуется и самим собой, **словно бы играет** «возвращение блудного сына», словно кто-то его еще со стороны наблюдает…*
[Андрей Битов. Заповедник (телемелодрама) (1991)]

Кроме того, в редких случаях позволяет использование настоящего или будущего времени союз *как если бы*.

---

[2] Здесь и далее используются примеры из Национального корпуса русского языка (www.ruscorpora.ru)

(4) *Части его в начале месяца распадаются и затем сходятся,* ***как если бы*** *на него* ***брызгают*** *сначала мертвой водой, а затем живой.* [Андрей Балдин. Московские праздные дни (1997)]

Примеры союза *как если бы* с настоящим временем единичны и периферийны, но представляют интерес. Союз комбинирует лексические элементы условия (*если*), сравнения (*как*) и ирреальности (*бы*). Сочетание *если бы* создает значение ирреального условия, которое блокирует возможность употребления изъявительного наклонения; сравнение, напротив, благоприятствует этому и создает возможность периферийного употребления изъявительного наклонения.

Остальные союзы допускают употребление лишь тех форм, которые характерны для сослагательного наклонения.

(5) *Никогда не слышала,* ***чтобы*** *он* ***звал*** *ее как-то иначе, чем «Рухэлэ».* [Дина Рубина. Окна (2011)]

(6) *Всем семьям просто необходимы как минимум три ребенка,* ***дабы не наступил*** *демографический кризис!* [«Русский репортер», № 18 (18), 2007]

(7) *Виктор Астафьев писал:* ***если бы*** *миллионы крестьян только* ***плюнули*** *в сторону Москвы, её бы смыло вместе с Кремлём и горийской обезьяной.* [Денис Драгунский. О рабах и свободных // «Частный корреспондент», 2011]

(8) *Пейзажи Соловков были бы еще живописнее, если это возможно,* ***когда бы*** *тут* ***было*** *побольше стад и птиц.* [В. И. Немирович-Данченко. Соловки (1874)]

(9) ***Кабы*** *все* ***носили*** *вещи так долго, не надо было бы создавать новые текстильные фабрики.* [Александр Чудаков. Ложится мгла на старые ступени (1987–2000) // «Знамя», 2000]

(10) *Время от времени я поглядывал на чайник, потому что* ***боялся, как бы*** *из носика не выплеснулось вино.* [Фазиль Искандер. Путь из варяг в греки (1990)]

(11) ***Пускай бы*** *все мы* ***погибли*** *до единого, но до той поры мы навели бы ужас на всю Германию своей дьявольской дерзостью и беспощадностью.* [А. И. Куприн. Последние рыцари (1934)]

(12) *И к тому же,* ***добро бы пострадал*** *Степанов; но ведь он же был оправдан обществом офицеров и без того.* [Ф. М. Достоевский. Подросток (1875)]

(13) *И* ***ладно бы тешились*** *только частные собственники, само градоначальство перекраивало город на лубочный лад.* [«Мир & Дом. City», 2003.10.15]

(14) *А ведь Мухаммед учил, что* […] *хозяин должен прощать ему проступки,* ***хотя бы*** *он* ***совершал*** *их по семьдесят раз на дню.* [«Наука и жизнь», 2008]

(15) …*Чиновники и государство сами берут на себя все расходы,* ***только бы*** *вы* ***начали*** *делать что-либо полезное обществу!* [Артем Тарасов. Миллионер (2004)]

(16) *Очередность особой роли не играет,* ***лишь бы*** *все нижеследующие этапы* ***были*** *завершены.* [коллективный. Отель «У погибшего альпиниста» (2008–2010)]

Итак, союзы *будто бы*, *как будто бы* и *словно бы* способны сочетаться с формами непрошедшего времени. Свидетельствует ли это однозначно в пользу того, что *бы* в этих сочетаниях не является показателем сослагательного наклонения? Нельзя исключить, что в том случае, когда сравнительные союзы сочетаются с прошедшим временем (с так называемой формой на *-л*), они содержат частицу сослагательного наклонения, а в остальных случаях сочетаются с изъявительным, то есть в примере (а) мы наблюдаем изъявительное наклонение, а в примере (б) — сослагательное:

(17) а) *Никто не смущался,* ***как будто бы*** *речь* ***идет*** *о чем-то заурядном.*
б) *Никто не смущался,* ***как будто бы*** *речь* ***шла*** *о чем-то заурядном.*

Тест, описанный в следующем разделе, позволяет доказать, что и в примере (17а), и в примере (17б) мы имеем дело с изъявительным наклонением, а частица *бы* является частью союза.

2.2. Возможность повторения частицы *бы*

Второй критерий, позволяющий признать частицу *бы* показателем сослагательного наклонения, — способность частицы повторяться, если есть сочиненный предикат.

Этот тест показывает, что форма прошедшего времени в сфере действия союзов *будто бы*, *как будто бы* и *словно бы* не является частью сослагательного наклонения, так как частица *бы* не может быть повторена:

(18) *В нашей комнате зверски пахло хлорофосом,* ***как будто бы*** *его щедрой рукой* ***прыскали*** *повсюду и потом* ***распахнули [*бы]*** *настежь все окна.* [Людмила Петрушевская. Тайна дома (1995)]

(19) ***Пила, будто бы*** *никогда* ***не читала*** *газет, и* ***не слушала [*бы]*** *радио, и не видела телевидения!* [Евгений Попов. «Пять песен о водке» (1970–2000)]

(20) *Горн издала странный звук,* ***словно бы*** *у шкафа* ***открылась и закачалась*** ***[*бы]*** *скрипучая дверца.* [Михаил Елизаров. Библиотекарь (2007)]

Все остальные союзы позволяют повторение частицы *бы*, если в сфере действия союза есть второе сказуемое:

(21) *Дмитрий Павлович, когда он за рулем, предпочитает,* ***чтобы звучало*** *радио и не* ***звучала бы*** *Тамара.* [Олег Зайончковский. Счастье возможно: роман нашего времени (2008)]

(22) *Пишу наскоро,* ***дабы*** *сегодня же* ***пошло*** *письмо и* ***застало бы*** *Вас в Вене.* [П. И. Чайковский. Переписка с Н. Ф. фон-Мекк (1883)]

(23) *И все равно,* ***если бы*** *я даже* ***поскользнулась*** *и* ***упала бы*** *перед всем классом, я сказала бы «An!».* [Алексей Беляков. Алка, Аллочка, Алла Борисовна (1998)]

(24) *И теперь мы далеки друг от друга,* ***как если бы*** *он* ***уехал*** *на Северный полюс и* ***остался бы*** *там на веки вечные.*

(25) *…Вы спрашиваете,* ***кабы прошло*** *время, и страсти* ***поутихли бы****, и она вернулась бы с повинной, стал бы я ее преследовать, казнить и все такое?* [Булат Окуджава. Путешествие дилетантов (1971–1977)]

(26) *…Графиня Клодина, видимо* ***опасаясь, как бы*** *ее подруга* ***не раздумала*** *и* ***не разрушила бы*** *этим весь хитро придуманный план, являлась к ней ежедневно и под каким-нибудь предлогом увозила ее из дому.* [Н. Э. Гейнце. Князь Тавриды (1898)]

(27) *Оправданы и даже совершенно необходимы разные подходы,* ***лишь бы*** *они* ***были*** *серьезными и* ***раскрывали бы*** *что-то новое в изучаемом явлении литературы…* [М. М. Бахтин. Ответ на вопрос редакции «Нового Мира» (1970)]

(28) *После войны она давала по два, а то и по три концерта в день, несмотря ни на какие трудности,* ***только бы*** *власти* ***разрешили*** *ее концерт,* ***дали бы*** *ей заработать.* [«Спецназ России», 2003.05.15]

(29) *…Он должен руководствоваться не своей собственной, а чужой, романо-германской национальной психологией* […]*,* ***хотя бы*** *это* ***противоречило*** *его национальной психологии, плохо* ***укладывалось бы*** *в его сознании.* [Н. С. Трубецкой. Европа и человечество (1920)]

В корпусе не нашлось таких примеров с союзами *ладно бы, добро бы, пусть бы*. Они нечастотны, и для них нехарактерны однородные сказуемые. Представляется, что следующие примеры с частицей *бы* при втором сказуемом допустимы:

(30) ***Ладно / добро / пусть бы*** *она* ***плакала*** *или* ***кричала бы****, а то весь день молчала.*

Таким образом, результаты теста на возможность повторения частицы при однородных сказуемых показывают такой же результат, что и тест на дистрибуцию форм при союзе: сравнительные союзы *будто бы, как будто* и *словно бы* не содержат показателя сослагательного наклонения (что не исключает семантической близости этих и некоторых других единиц с зоной ирреальности). Что касается остальных рассмотренных союзов, то глагольная форма, которая находится в их сфере действия, должна вместе с частицей *бы* считаться формой сослагательного наклонения, а сам союз в строгом смысле является сочетанием союза и грамматической частицы *бы*.

Отметим, что повтор частицы возможен и тогда, когда в предложении лишь одно сказуемое:

(31) *Честно говоря, очень хочется,* ***чтобы*** *энергии Магомеда Толбоева обязательно* ***хватило бы****, чтобы одолеть и эту проблему.* [«Наука и жизнь», 2009]

Такие примеры являются свидетельством обратного процесса: союз *с* частицей *бы* не воспринимается говорящим как носитель наклонения, и потому говорящий добавляет еще одну частицу. Вероятно, разные союзы в разной степени допускают такой повтор (см. об этом Летучий 2014). Проверка союзов на частотность избыточного употребления частицы, возможно, позволила бы добавить еще один критерий для иерархии союзов с точки зрения того, насколько частица *бы* инкорпорирована в союз. В настоящей работе такая проверка не производилась.

3. Союзы с частицей *бы*: два слова или одно

Ричардом Брехтом было высказано мнение, что союз *чтобы* можно считать орфографическим вариантом сочетания *что* и *бы* (Brecht 1977: 38). Чтобы проверить степень отделимости элемента *бы*, нужно ответить на следующие вопросы:

- Может ли частица *бы (б)* быть опущена, так чтобы значение придаточного изменилось только за счет изменения наклонения?
- Может ли частица *бы (б)* находиться дистантно по отношению к союзу без изменения значения конструкции?

Эти вопросы не могут быть заданы по отношению к словам *дабы* и *кабы*, поскольку в современном языке нет союзов *ка* и *да*.

3.1. Опущение частицы

С точки зрения возможности опустить частицу *бы (б)* союзы различаются.

К союзным сочетаниям, которые позволяют опущение, то есть имеют соответствие в индикативе, относятся следующие: *словно бы, будто бы, как будто бы, если бы, как если бы, добро бы, ладно бы, пускай бы, хоть бы, хотя бы*,

а также союз *чтобы* в некоторых своих употреблениях. Ср. примеры (1), (2), (3), (7) и следующие:

(32) *Сложно поверить,* ***будто*** *она этого не знает.*

(33) *Она так и осталась буквой «г»,* ***как будто*** *она тянет весь груз своей жизни.*

(34) *Он любуется самим собой,* ***словно*** *играет возвращение блудного сына.*

(35) ***Если*** *миллионы крестьян только* ***плюнут*** *в сторону Москвы, её смоет вместе с Кремлём и горийской обезьяной.*

Сравнительный союз *как если бы* имеет индикативную пару, хотя такие конструкции нечастотны:

(36) *Плюгавого каптерщика Матюшин невзлюбил еще за старое, но не родилось в нем злости,* ***как если попался*** *по пути не человек, а гриб-гнилушка.* [Олег Павлов. Дело Матюшина (1996)]

Для уступительных союзов *хотя, хоть* и *пускай* индикативный вариант является даже более частотным, чем с частицей *бы*:

(37) а) *Хозяин простит слуге проступки,* ***хотя бы*** *он* ***совершал*** *их по семьдесят раз на дню.*
б) *Хозяин прощал слуге проступки,* ***хотя*** *он* ***совершал*** *их по семьдесят раз на дню.*

(38) а) ***Хоть бы*** *и* ***слышал****, все равно бы не понял.*
б) ***Хоть*** *и* ***слышал****, все равно не понял.*

(39) а) ***Пускай бы*** *все* ***погибли****, но их жизнь прошла бы не зря.*
б) ***Пускай*** *все* ***погибли****, но их жизнь прошла не зря.*

Примеры *ладно* в значении уступительного союза без частицы *бы* редки, но возможны.

(40) ***Ладно*** *привязать охвостку* ***догадался*** *за камень на берегу, иначе не нашел бы мережи. Попробовал подтянуть сеть с плота — она не сдвинулась с места.* [Виктор Астафьев. Царь-рыба (1974)]

Для союза *добро бы* контекстов без частицы *бы* найти не удалось. По аналогии с союзом *ладно бы* можно предположить, что они возможны, но конструирование таких примеров затруднено тем, что союз является устаревшим и интуитивно не всегда очевидна возможность тех или иных употреблений:

(41) ?***Добро пострадали*** *только военные, но ведь могли и мирные жители.*

Союз *чтобы* весьма частотен и разнообразен в своих употреблениях. Некоторые конструкции имеют индикативную пару (42 и 43), большая часть существует только с сослагательным наклонением (44). Объемы настоящей статьи не позволяют рассмотреть все случаи.

(42) а) *Не думаю,* ***чтобы*** *твоя мама* ***была*** *в восторге от столь раннего брака.*
б) *Не думаю,* ***что*** *твоя мама* ***будет*** *в восторге от столь раннего брака.*

(43) а) *Светильник поставили* ***так, чтобы*** *его увидело максимальное число людей.*
б) *Светильник поставили* ***так, что*** *его увидело максимальное число людей.*

(44) а) *Хочу,* ***чтобы*** *меня выслушали.*
б) **Хочу,* ***что*** *меня выслушали / выслушают.*

(45) а) *Пришел,* ***чтобы*** *ему помогли.*
б) **Пришел,* ***что*** *ему помогли / помогут.*

Следующие союзы не имеют индикативных пар: *как бы, когда бы, кабы, дабы, только бы, лишь бы* и союз *чтобы* в части своих употреблений. Например:

(46) а) *Он мечтал,* ***как бы выследить*** *тигра.*
б) **Он мечтал,* ***как выследить*** *тигра.*

(47) а) *Она боялась,* ***как бы не*** *опоздать.*
б) **Она боялась,* ***как не*** *опоздать.*

Представляется невозможным опущение частицы в союзе *когда бы*. Хотя в ряде грамматических справочников приводятся примеры условного употребления союза *когда*, все они не очень убедительны. Русская грамматика рассматривает как имеющий условное значение (РГ 1980: 563–568) только союз *когда бы*.

Не удалось найти индикативные пары к конструкциям с оптативно-уступительными союзами *только бы* и *лишь бы*.

Итак, с точки зрения возможности опущения частицы *бы (б)* союзы разделились на две группы: те, которые имеют варианты без частицы *бы*, и те, которые таких вариантов не имеют. Это свойство союзов может быть сопоставлено с другим — с возможностью дистантной позиции частицы *бы (б)* по отношению к союзу.

3.2. Дистантное положение частицы

Вопрос о возможности дистантного расположения частицы *бы* по отношению к союзу в ряде случаев не может быть решен однозначно, потому что требует семантического анализа: если примеры с дистантным расположением частицы

существуют, можно ли их считать полностью аналогичными примерам с контактным расположением? Здесь будут рассмотрены лишь наиболее очевидные случаи.

Союзы *если бы*, *как если бы, добро бы, ладно бы, пускай бы* допускают дистантное употребление (в работе Bonnot & Bottineau 2011 высказаны некоторые предположения относительно того, с чем связано место частицы в условных придаточных предложениях):

(48) ***Если*** *миллионы крестьян только* ***плюнули бы*** *в сторону Москвы, её бы смыло вместе с Кремлём и горийской обезьяной.*

(49) *Так же искренне,* ***как если сказал бы****, что из фруктов мне нравятся яблоки.* [Нодар Джин. Учитель (1980–1998)]

(50) ***Ладно были бы*** *новобранцы, а то ведь прекрасно знают, почем фунт лиха, и все равно лезут.* [Андрей Белянин. Свирепый ландграф (1999)]

(51) ***Добро были бы*** *обязаны, а то жертвуете собою из любви к искусству!* [В. В. Верещагин. Литератор (1894)]

(52) ***Пускай были бы*** *новобранцы, а то ведь опытные солдаты.*

Трудно решить вопрос о возможности дистантного расположения *бы* при союзе *когда*, поскольку союз является устаревшим:

(53) ?*Пейзажи Соловков были бы еще живописнее,* ***когда было бы*** *побольше стад и птиц.*

Союзы *лишь бы, только бы* и *хоть бы* допускают дистантно употребление, хотя такие примеры весьма редки:

(54) *Вряд ли сейчас можно сомневаться и в том, что творчеству, как и всему другому, можно научиться,* ***лишь были бы*** *осуществлены необходимые условия.* [«Химия и жизнь», 1967]

(55) *И сил у меня еще бы хватило, чтобы запустить механизмы прогресса,* ***только были бы*** *воля, понимание и поддержка власти.* [Артем Тарасов. Миллионер (2004)]

(56) *Если больше страницы — читать не станут,* ***хоть*** *это* ***был бы*** *шедевр.* [Самуил Алешин. Встречи на грешной земле (2001)]

Союз *хотя бы*, по-видимому, нет:

(57) **Хозяин простит слуге проступки,* ***хотя*** *он* ***совершал бы*** *их по семьдесят раз на дню.*

Что касается союза *чтобы*, то возможность дистантного расположения частицы *бы* сильно различается для разных типов союза. Есть примеры, где союз *чтобы* и сочетание *что* + *бы* полностью синонимичны. Все эти случаи относятся к типу эпистемических придаточных предложений, но обратное неверно — не все эпистемические придаточные допускают дистанцию (о различии между эпистемическим и целевым типом см. Dobrushina 2012):

(58) а) *Не думаю,* ***чтобы*** *твоя мама* ***была*** *в восторге от столь раннего брака.*
б) *Не думаю,* ***что*** *твоя мама* ***была бы*** *в восторге от столь раннего брака.*

Для союза *чтобы* целевого типа дистантных употреблений не обнаруживается, даже если рассматривать те придаточные, для которых характерна конкуренция между сослагательным наклонением и индикативом (в работе Spencer & Luís 2012: 217 на этом основании предложено считать *бы* в составе этого союза морфемой):

(59) а) *Как мог, я* ***способствовал тому, чтобы*** *роман вышел в свет.*
[Г. Я. Бакланов. Жизнь, подаренная дважды (1999)]
б) *Как мог, я* ***способствовал тому, что*** *роман выйдет в свет.*
в) **Как мог, я* ***способствовал тому, что*** *роман вышел* ***бы*** *в свет.*

Препятствием для такого разделения *что* и *бы* является то, что в независимом употреблении частица *бы* создает контрфактивное значение, неуместное в этом предложении.

Не допускают отделения частицы союзы *будто бы, как будто бы* и *словно бы:*

(60) **Сложно поверить,* ***будто*** *она этого не знает* ***бы****.*

(61) **Она так и осталась буквой «г»,* ***как будто*** *она тянет* ***бы*** *весь груз своей жизни.*

(62) **Он любуется самим собой,* ***словно*** *играет* ***бы*** *возвращение блудного сына.*

Критерий отделимости требует и обратной процедуры: анализа тех примеров, где *бы* находится на дистанции от союза, с точки зрения возможности переноса частицы в постпозицию к союзу. Полный обзор таких случаев здесь невозможен, приведем лишь несколько примеров, когда перенос недопустим.

(63) *Он рассердился, быстро закончил заседание, а наши сотрудницы потом говорили,* ***что*** *они* ***бы не решились*** *на открытое выступление.* [«Даша», 2004]

(64) *У них такие дипломы,* ***что*** *во Франции они* ***получали бы*** *с ним, ну, тысячу долларов.* [«Столица», 1997.06.17]

(65) *В суд подавать не стали,* ***хотя*** *точно выиграли* ***бы*** *дело.* [«Русский репортер», № 3 (181), 27 января 2011, 2011]

4. Заключение

Было показано, что союзы с частицей *бы (б)* неоднородны с точки зрения того, можно ли считать их сочетанием союза и частицы сослагательного наклонения, и с точки зрения того, насколько расчленимым является это сочетание. Более того, как видно из таблицы 1, расчленимость является градуальным свойством, поскольку третий и четвертый критерии могут давать разные результаты.

Выяснилось, что критерий опущения частицы не коррелирует с тем, является ли частица показателем сослагательного наклонения: сравнительные союзы *будто бы, как будто бы* и *словно бы* могут употребляться без частицы.

Таблица 1. Свойства союзов с частицей *бы*

	дистрибуция форм	**повторяемость**	**опущение**	**дистантная позиция**
чтобы	сосл.	да	да / нет	да / нет
дабы	сосл.	да	нет	нет
если бы	сосл.	да	да	да
когда бы	сосл.	да	нет	?
кабы	сосл.	да	нет	нет
как бы	сосл.	да	нет	нет
будто бы	индик.	нет	да	нет
как будто бы	индик.	нет	да	нет
словно бы	индик.	нет	да	нет
как если бы	сосл.	да	да	да
добро бы	сосл.	да	да?	да
ладно бы	сосл.	да	да	да
лишь бы	сосл.	да	нет	да
только бы	сосл.	да	нет	да
хотя бы	сосл.	да	да	нет
хоть бы	сосл.	да	да	да
пускай бы	сосл.	да	да	да

Литература

1. *Летучий А. Б.* (2014). Сентенциальные актанты. Материалы для проекта корпусного описания русской грамматики (http://rusgram.ru). На правах рукописи. М.
2. *Апресян В. Ю., Пекелис О. Е.* (2012) Союз. Материалы для проекта корпусного описания русской грамматики (http://rusgram.ru). На правах рукописи. М.

3. *Русская грамматика*. (1980). Под ред. Н. Ю. Шведовой. Издательство «Наука». Т. II.
4. *Bonnot, C., & Bottineau, T.* (2013). « Lorsque la marque du conditionnel est une particule mobile : le cas du russe », in Ultériorité dans le passé : le conditionnel, Direction J. Brès, S. Sarrazin, S. Azzopardi, Faits de langue, № 40, pp. 189–196.
5. *Brecht, R. D.* (1977). Čtoby or čto and by. Folia Slavica. Columbus, Ohio, 1(1), 33–41.
6. *Dobrushina Nina.* (2012). Subjunctive complement clauses in Russian. Russian Linguistics, 36(2).
7. *Spencer, Andrew, and Ana R. Luís.* (2012) Clitics: an introduction. Cambridge University Press.

References

1. *Letuchij A. B.* (2014). Sentencial'nye aktanty. Materialy dlja proekta korpusnogo opisanija russkoj grammatiki (http://rusgram.ru). [Complement clauses. Towards a corpus description of Russian grammar]. Manuscript. M., 2014.
2. *Apresjan V. Ju., Pekelis O. E.* (2012) Conjunction. Materialy dlja proekta korpusnogo opisanija russkoj grammatiki (http://rusgram.ru) [Parts of speech. Towards a corpus description of Russian grammar]. Manuscript. M.
3. *Russkaja grammatika.* [Russian grammar]. (1980). Pod red. N.Ju.Shvedovoj. Izdatel'stvo «Nauka». T. II.
4. *Bonnot, C., & Bottineau, T.* (2013). « Lorsque la marque du conditionnel est une particule mobile : le cas du russe », in Ultériorité dans le passé: le conditionnel, Direction J. Brès, S. Sarrazin, S. Azzopardi, Faits de langue, № 40, pp. 189–196.
5. *Brecht, R. D.* (1977). Čtoby or čto and by. Folia Slavica. Columbus, Ohio, 1(1), 33–41.
6. *Dobrushina Nina.* (2012). Subjunctive complement clauses in Russian. Russian Linguistics, 36(2).
7. *Spencer, Andrew, and Ana R. Luís.* (2012) Clitics: an introduction. Cambridge University Press.

ИНТРОДУКЦИЯ РЕФЕРЕНТА В РУССКИХ УСТНЫХ ПЕРЕСКАЗАХ (НА МАТЕРИАЛЕ «РАССКАЗОВ О ГРУШАХ» У. ЧЕЙФА)[1]

Федорова О. В. (olga.fedorova@msu.ru)

МГУ имени М.В. Ломоносова и Институт языкознания РАН, Москва, Россия

В серии работ, опубликованных двадцать лет назад на материале анализа русских, немецких и шанских[2] сказок, мы исследовали вопрос о типологии средств интродукции референта в письменных текстах. Цель настоящей работы состояла в том, чтобы на материале устных пересказов известного «Фильма о грушах» Уоллеса Чейфа оценить, насколько разработанная «сказочная» модель интродукции применима к устному дискурсу; в качестве альтернативного подхода была рассмотрена модель Чейфа, разработанная на материале английских пересказов «Фильма о грушах». Настоящее исследование было выполнено на материале 25 русских пересказов, проанализированный корпус включает 125 интродуктивных предложений. Предлагаемая модель интродукции отличается как от «сказочной» модели интродукции, так и от модели Чейфа по каждому из пяти рассмотренных пунктов: тип привязки, наличие речевых сбоев, статус персонажа и распространенность предложения, ограничение на легкое подлежащее, ограничение одного нового понятия. Выявленные тенденции должны быть подтверждены на более обширном материале. Кроме того, некоторые из обнаруженных закономерностей ждут продолжения исследования на материале других, неинтродуктивных, предложений русского языка.

Ключевые слова: интродукция, устный дискурс, «Рассказы о грушах», русский язык

REFERENT INTRODUCTION IN RUSSIAN SPOKEN NARRATIVES

Fedorova O. V. (olga.fedorova@msu.ru)

Lomonosov Moscow State University, Moscow, Russia

[1] Работа выполнена при финансовой поддержке РФФИ (проект № 14-06-00211). Автор выражает благодарность А. А. Кибрику и трем анонимным рецензентам за критические замечания, высказанные при подготовке работы.

[2] Шанский язык — один из тайских языков, распространен в основном в Мьянме и Китае.

In a series of papers published twenty years ago on analysis of Russian, German and Shan tales, we examined the typology of referent introduction in written texts. The general purpose of the current study was to evaluate how the “tale” model of introduction is applicable to the spoken narratives; as an alternative approach we considered the Chafe’s model, based on the English “Pear stories” (Chafe 1980). Twenty five Russian participants took part in the experiment; all the participants described the same experimental film about some child stealing pears; thus we analyzed 25 narratives and 125 introductive sentences. Surprisingly, our model differs from both the “tale” introductive model and the Chafe’s model for each of the following points: (1) type of the common ground, (2) speech disfluencies, (3) the character status and clauses number, (4) the “light subject” constraint, (5) the “one new idea” constraint. However, all of these results need further empirical justification in new studies on Russian materials.

Key words: introduction, spoken discourse, “Pear stories”, Russian language

1. «Сказочная» интродуктивная модель

В основу данного исследования положены исходные представления о феномене интродукции, введенные в (Федорова 1994); под интродукцией мы понимаем введение нового референта в долговременную память адресата. В прототипическом случае интродуктивное предложение состоит из трех основных элементов: привязка — бытийный оператор — номинация референта. ***Привязкой*** мы назвали общие знания о фрагменте действительности, которые, по мнению говорящего, имеются и у говорящего и у адресата; более употребительный современный эквивалент — ***общая позиция*** (common ground), в данной работе мы сохраним термин «привязка». В работе (Федорова 1994) на материале текстов сказок был обоснован постулат о принципиальной небходимости привязки; были выделены эксплицитные и имплицитные привязки, а также ступенчатая привязка и псевдопривязка — введение героя повествования через второстепенных или эпизодических персонажей, особенно частотно это явление в абсолютном начале сказочных текстов, напр.: *Жили старичок со старушкою; у них была дочка да сынок маленький* (Гуси-лебеди). ***Бытийный оператор*** указывает на интродуктивную неопределенную референцию имени. Номинация состоит из ***интродуктива*** (который во многих европейских языках обычно выражается неопределенным артиклем) и собственно номинации.

В (Федорова 1994) было показано, что при интродукции референта говорящий может использовать различные языковые средства: порядок слов, отдельное интродуктивное предложение, употребление имени собственного и др. В разных языках проблема выбора набора средств, отвечающих за введение референта, решается по-разному; в некоторых удается выделить какое-либо доминирующее средство (в частности, «говорящие имена» в шанском языке), в других интродуктивный потенциал оказывается равномерно распределен между многими способами кодирования. Еще один важный аспект связан с понятием ***силы интродукции***, то есть прагматической потребностью в ней.

На материале сказок было показано, что чем выше место персонажа в иерархии «главный герой — герой — второстепенный персонаж — эпизодический персонаж», тем сильнее интродукция.

Цель настоящей работы состоит в том, чтобы оценить, насколько разработанная на материале письменных сказок модель интродукции применима к устному дискурсу и при необходимости скорректировать ее. В качестве устного материала были выбраны известные «Рассказы о грушах» У. Чейфа (Chafe 1980). Видеоролик, созданный авторами специально для этого научного проекта, был выбран по многим причинам; в частности, на основании разработанных в (Федорова 1994) критериев в нем однозначно выделяются: главный герой (мальчик), герой (садовник), второстепенный персонаж (три мальчика), а также эпизодические персонажи (мужчина с козой и девочка). Опишем краткий сюжет фильма:

(1) Мужчина на дереве собирает груши. Мимо проходит мужчина с козой. Приезжает мальчик на велосипеде, берет одну корзину с грушами и уезжает. По дороге навстречу мальчику едет девочка на велосипеде; у мальчика слетает шляпа; велосипед наезжает на камень; мальчик падает; груши рассыпаются. Три подошедших мальчика помогают собрать груши и отдают потерянную шляпу. Мальчик дает им три груши; они уходят в разные стороны. Мужчина на дереве продолжает собирать груши; он обнаруживает пропажу одной корзины с грушами. Мимо проходят три мальчика, жуя подаренные груши. Мужчина смотрит им вслед.

Если следовать канонической сказочной модели интродукции, введение персонажей будет выглядеть примерно так:

(2) [садовник] В одной мексиканской деревне жила-была одна семья, у которой был большой грушевый сад. Однажды солнечным летним днем глава семьи — крупный усатый мужчина средних лет в шляпе, белом фартуке и с красным платком на шее — отправился в сад собирать груши. <…> [мужчина с козой] В этот момент мимо дерева проходит мужчина, который ведет козу. <…> [мальчик] Тут к грушевому дереву на большом красном велосипеде старой модели подъезжает десятилетний мальчик в широкой шляпе, на шее у мальчика тоже платок. <…> [девочка] Навстречу мальчику по дороге едет девочка на велосипеде. <…> [три мальчика] К мальчику подходят трое ребят и помогают ему подняться, один из мальчиков играет пинг-понговым шариком.

Даже при беглом взгляде на (2) кажется очевидным, что введение персонажей при устных пересказах не происходит подобным образом. В частности, плавность речи нарушается паузами хезитации и другими речевыми сбоями, сила интродукции не так однозначно связана с ролью персонажа, а характерная для сказок псевдопривязка заменяется какими-то иными средствами интродукции. Прежде чем приступить к анализу материала русских пересказов, обратимся

к исследованиям У. Чейфа, выполненным на материале английских пересказов «Фильма о грушах». Мы ожидаем, что модель Чейфа окажется более применима к нашему материалу, чем модель, основанная на письменном сказочном дискурсе.

2. Интродукция персонажей по Чейфу

В данном разделе мы рассмотрим два дискурсивных ограничения из (Chafe 1994), используемых при введении новых персонажей — ограничение на «легкое подлежащее» (light subject constraint) и ограничение одного нового понятия (ООНП, one new idea constraint).

Как известно, Чейф выделяет три когнитивных статуса референта в сознании (consciousness) адресата: активный, полуактивный и неактивный. Активную информацию говорящий обычно упаковывает как данную (given), неактивную как новую (new), а полуактивную — как доступную (accessible); доступные референты находятся в полуактивном состоянии благодаря предыдущим упоминаниям.

Чейф отмечает, что предложения *A girl saw John* ‘Какая-то девушка увидела Джона’, регулярно встречающиеся в научных публикациях, редко употребляются в реальной речи. Проанализировав разговорный корпус английского языка, состоящий из 10 тыс. словоупотреблений, он пришел к выводу, что 81 % всех подлежащих вводят данную информацию, причем 98 % из них являются местоимениями (Chafe 1994: 85). Оставшиеся 19 % подлежащих в 16 % вводят доступную информацию и только в 3 % новую; в последних случаях таким образом вводится тривиальная информация, которая никогда больше не упоминается (Chafe 1994: 88–91). Таким образом, ограничение на «легкое подлежащее» в разговорной речи гласит, что подлежащее вводит или неновую информацию, или (если новую) несущественную информацию (Chafe 1994: 92).

Второе ограничение, которое описывает Чейф, касается количества новой информации, вводимой в одной ***интонационной единице*** (ИЕ, intonation unit). ИЕ по Чейфу отражает текущий фокус сознания и часто совпадает с клаузой (clause); в данном исследовании мы будем использовать близкий термин ***элементарная дискурсивная единица*** (ЭДЕ, см. (Кибрик, Подлесская (ред.) 2009). Анализируя пример *Jennifer was really happy* ‘Дженифер действительно счастлива’, Чейф пишет, что данное высказывание может встретиться в реальной речи только в том случае, если концепт Дженифер активирован в сознании адресата (Chafe 1994: 108). Таким образом, прототипическая ИЕ состоит из легкого подлежащего, в котором содержится данная информация, и предиката, содержащего новую информацию (Chafe 1994: 108). Анализ Чейфа основан на собранных пересказах «Фильма о грушах», ниже приводится интродукция персонажей одного из английских пересказов в переводе на русский язык[3]:

---

[3] Русский перевод лишь условно соответствует английскому оригиналу; в частности, в оригинале часто используется конструкция There is. В скобках здесь и далее указывается длительность пауз.

(3) [садовник] (.35) Вот тут (.65) какой-то фермер, (.15) он выглядит как ээ . . американский мексиканец, (.5) он собирает груши. <...> [мужчина с козой] Ээ . . и какой-то мм . . какой-то мужчина с козой (.2) проходит мимо. <...> [мальчик] (.4) И маленький мальчик, он приезжает на своем велосипеде. <...> <...> [девочка] И вот тут еще какая-то девочка, (.35) едет на велосипеде, примерно его возраста, (.25) по дороге. <...> [три мальчика] Между тем вот тут три маленьких мальчика, (.15) недалеко на дороге.

Итак, мы рассмотрели два примера интродукции персонажей при пересказе «Фильма о грушах» — один пример был сконструирован на основании сказочной модели интродукции, второй взят из реальных «Рассказов о грушах» и переведен на русский язык. В следующем разделе мы оценим, насколько реальные русские пересказы «Фильма о грушах» соответствуют первой или второй модели.

3. Интродуктивная модель в русских «Рассказах о грушах»

Данное исследование выполнено на материале 25 пересказов «Фильма о грушах», в качестве испытуемых были привлечены студенты и сотрудники МГУ имени М.В. Ломоносова в возрасте от 17 до 46 лет. Исследование проводилось индивидуально: сначала каждый испытуемый смотрел шестиминутный видеоролик, который он раньше не видел, а затем пересказывал его содержание второму участнику. Пересказы были записаны на видеокамеру и диктофон, расшифрованы и затранскрибированы; деление на предложения и ЭДЕ было произведено на основании просодических критериев; в программе PRAAT (www.fon.hum.uva.nl/praat) была размечена временная динамика и выделены паузы. В данной работе проанализированы фрагменты пересказов, в которых вводятся основные персонажи данного фильма. Таким образом, корпус интродуктивных предложений содержит 125 единиц, по 25 для каждого из пяти персонажей.

Собранные интродуктивные предложения были проанализированы по следующим основаниям: тип привязки (раздел 3.1), наличие речевых сбоев (3.2), статус персонажа и распространенность предложения (3.3), ограничение на легкое подлежащее (3.4), ограничение одного нового понятия (3.5).

3.1. Тип привязки

В сказочных интродукциях в абсолютном начале используются псевдопривязка и ступенчатая привязка; при интродукции важных персонажей обычно используется привязка по времени / месту, при интродукции неважных — привязка через другие персонажи. В английских «Рассказах о грушах», согласно транскриптам из (Chafe (ed.) 1980) и работе (DuBois 1980), во всех случаях доминирует использование привязки по времени / по месту. Посмотрим, какие типы привязки встречаются в русских пересказах.

Как и ожидалось, псевдопривязка и ступенчатая привязка не характерны для пересказов, такие конструкции встретились в нашем корпусе только по одному разу, см. пример ***ступенчатой*** привязки[4]:

(4) показывается дерево,
(.4) и к нему присоединена лестница,
ээ(.3) на этой лестнице мужчина.

Однако, в отличие от английских, для русских пересказов оказалась характерна так называемая ***кинематографическая*** привязка, т. е. использование в качестве общего знания того факта, что рассказчик пересказывает сюжет фильма. Такая привязка хотя бы раз встретилась в каждом из 25 пересказов. Например, интродукция мужчины с козой выглядит так:

(5) ээ(.2) потом мы слышим (.3) звук блеяния козы,
(.5) видим человека,
(.4) который ведет козу на веревке.

Вопрос использования кинематографического взгляда в «Рассказах о грушах» был впервые описан в (Tannen 1980). Оказалось, что американские испытуемые чаще пересказывали видеоролик как фильм, в то время как греческие испытуемые просто рассказывали историю, не упоминая о том, что действие происходит в фильме; согласно работе (Mazur and Chmiel 2012) в большинстве «Рассказов о грушах» кинематографическая лексика встречается хотя бы один раз; в частности, для голландских пересказов эта цифра составляет 79%, для греческих 80%, для польских 85%.

В наших пересказах кинематографический взгляд использовался хотя бы один раз в 100% пересказов. Более того, была выявлена важная закономерность: количество его использования для нужд интродукции последовательно сокращается от начала к концу пересказа независимо от статуса персонажа: садовник был введен при помощи кинематографической привязки в 23 случаях из 25, мужчина с козой в 10 случаях, мальчик в 9 случаях, девочка в 3 случаях, три мальчика в 1 случае.

Рассмотрим два другие типа привязок: по времени / месту и через уже введенного персонажа. Садовник 2 раза вводится при помощи привязки по времени / месту, мужчина с козой 3 раза вводится привязкой по времени / месту и 12 раз через садовника, мальчик 12 раз вводится привязкой по времени / месту и 4 раза через садовника, девочка ни разу не вводится привязкой по времени / месту и в 22 случаях вводится через мальчика, три мальчика 5 раз вводятся привязкой по времени / месту и 19 раз через мальчика. Таким образом, наблюдаемая закономерность совпадает с правилом сказочной интродукции: привязка по времени / месту оказывается характерна для важных персонажей, а привязка

[4] Здесь и далее примеры даются с разбивкой на ЭДЕ, символами ээ и мм обозначены заполненные паузы.

через другие персонажи — для неважных. Однако в отличие от сказочной интродукции доминирующим средством оказывается кинематографическая привязка.

3.2. Речевые сбои

Как мы и предполагали, речевые сбои (в частности, заполненные и незаполненные паузы хезитации) оказались самым ярким отличием устных пересказов от письменных сказочных текстов. На таком ограниченном материале мы можем сделать только самые предварительные выводы, которые нуждаются в более серьезном подтверждении. Тем не менее, по нашим данным, паузы хезитации при интродукции любого персонажа оказываются заметно длиннее, чем паузы в начале других эпизодов, в которых введения персонажей не происходит. Более того, эти длинные паузы часто встречаются внутри предложения, перед вложенной клаузой, что в общем случае представляет собой нехарактерное явление:

(6) ээ(.4) мальчик едет по дороге,
(.7) мм(.8) видит едущую навстречу ему (.2) девочку,

3.3. Статус персонажа и распространенность вводящего его предложения

В сказочных дискурсах важные персонажи вводятся с большим количеством дополнительной информации в том же предложении. При рассмотрении этого вопроса в (Федорова 1994) мы считали общее число значимых дополнительных признаков («возраст», «социальный статус» и проч.) без учета структуры предложения. На материале устных пересказов мы обратились к вопросу о том, есть ли связь между статусом персонажа и распространенностью вводящего его предложения. Сосчитав количество клауз для всех 125 интродуктивных предложений, мы получили такие усредненные цифры: при интродукции садовник а используется 3.2 клаузы, при интродукции мужчины с козой — 2.4, мальчика — 2.5, девочки — 3.1, трех мальчиков — 3.2. Как видно, данное распределение оказывается никак не связано со статусом персонажа, а одинаковые цифры при интродукции садовника и трех мальчиков говорят, по-видимому, о разных тенденциях. В первом случае речь идет о дополнительных интродуктивных средствах, необходимых при введении персонажа в абсолютном начале (аналог псевдопривязки в сказках); во втором случае работает противоположное правило: при интродукции неважных персонажей в одном предложении аккумулируется много информации, которая при введении более важных персонажей обычно упаковывается в несколько предложений, ср. (7) и (8):

(7) после этого (.2) вдруг появляются трое мальчиков,
(.5) ниоткуда,
(.3) двое постарше (.2) а один помладше,

(.3) и они ему помогают мм(.2) собирают,
ээ(.3) поднимают его,
(.3) собирают те груши,
(.4) кладут в корзину.

(8) (1.0) дальше камера наезжает на мальчика,
(.2) и мальчик на велосипеде.
(.4) велосипед ещё такой старый как раз,
(.7) с верхней рамой.
(2.0) ээ(.5) мальчик одет (.6),
я бы сказал как мм(.2) пионер,
ну довольно цивильно,
(.6) вот то есть он не шпана.

При более внимательном взгляде оказалось, что в многоклаузных предложениях типа (7) часто содержится информация не только о вводимом персонаже:

(9) ээ(.2) по пути он (.2) встретил девочку на велосипеде,
(.5) ээ(.2) столкнулся с ней,
(.2) с него слетела шляпа,
(.4) и он (.3) упал.

В последнем примере две из четырех клауз не имеют отношения к девочке. Если пересчитать данные, учитывая только клаузы с интродукцией данного персонажа, при интродукции садовник а используется 2.2 клаузы, при интродукции мужчины с козой — 2, мальчика — 2.2, девочки — 1.5, трех мальчиков — 2.5. Самое значительное уменьшение наблюдается в случае интродукции девочки, что хорошо согласуется с эпизодическим статусом персонажа.

Таким образом, статус персонажа никак не связан с распространенностью предложения, а неважные персонажи часто вводятся с большим количеством дополнительной информации.

3.4. Ограничение на легкое подлежащее

В (Федорова 1994) мы не затрагивали вопрос о легком подлежащем, а по работам Чейфа известно, что 81% подлежащих вводится как данное. Подсчитав синтаксическую роль персонажей, мы получили цифры, обратные статистике Чейфа: в 85,6% персонаж вводится именно в позиции подлежащего. Процент мог быть еще выше, если бы не характерная для девочки интродукция в позиции дополнения, см. пример 8.

Как видно, на нашем материале это ограничение Чейфа не работает. Возможно, это связано со свободным порядком слов в русском языке по сравнению с английским. Насколько это ограничение нарушается на русском материале

в неинтродуктивных предложениях в пересказах «Фильма о грушах», а также на другом русском материале, остается вопросом для дальнейшего изучения.

3.5. Ограничение одного нового понятия

Второе ограничение Чейфа — ООНП — оказалось распределено в наших 25 пересказах следующим образом. В случае интродукции садовника данный принцип не соблюдается в 20 случаях из 25, например:

(10) ээ(0,3) всё начинается с того,
что мужчина собирает груши.

Однако при введении других персонажей этот принцип не нарушается ни разу. Можно предположить, что в русских нарративах, в отличие от английских, позиция абсолютного начала повествования является более выделенной и требует использования особых дискурсивных средств, подчеркивающих этот статус. Это предположение также нуждается в проверке на более обширном материале.

4. Заключение

Проанализировав 125 русских интродуктивных предложений, взятых из 25 пересказов «Фильма о грушах» Чейфа, мы получили следующие результаты:

1. Привязка. Наиболее распространенной оказалась кинематографическая привязка, частота использования которой уменьшается к концу пересказа; использование других привязок коррелирует со статусом персонажа: привязка по времени / месту характерна для важных персонажей, а привязка через другие персонажи — для неважных.
2. Паузы хезитации, в целом характерные для устной речи, в случае интродукции становятся более длинными даже в нехарактерных для них местах в середине предложения.
3. Статус персонажа оказывается никак не связан с распространенностью вводящего его предложения, а неважные персонажи часто вводятся с большим количеством дополнительной информации.
4. В 85,6% новая информация о персонаже вводится в позиции подлежащего.
5. Ограничение одного нового понятия нарушается при интродукции садовника и соблюдается при интродукции других персонажей.

Типичный усредненный пересказ может выглядеть так:

(11) [садовник]
(.5) ээ(.5) фильм начинается с того что
мм(.6) усатый мужчина в шляпе собирает груши.
[мужчина с козой]

(.6) в это время ээ(.3) мимо проходит какой-то мужчина с козой,
коза упирается.
[мальчик]
(.8) ээ(.4) затем ээ(.2) приезжает маленький мальчик на велосипеде,
в большой шляпе.
[девочка]
(.4) он ээ(.2) едет через поле,
(.3) а навстречу ему едет девочка,
(.4) тоже на велосипеде.
[три мальчика]
(1.0) это видят (.5) трое других мальчиков,
(.2) мм(.3) они подходят и помогают ему подняться,
(.4) и собирают груши в корзину.

Описанная модель интродукции отличается как от «сказочной» модели интродукции (в которой средства интродукции выбираются в зависимости от важности персонажа), так и от модели Чейфа по каждому из пяти рассмотренных пунктов. Оказалось, что выбор интродуктивной стратегии в первую очередь зависит от позиции вводимого персонажа в тексте. Выявленные тенденции должны быть подтверждены на более обширном материале. Кроме того, некоторые обнаруженные закономерности (в частности, нарушение обоих ограничений Чейфа) ждут продолжения исследования на материале других, неинтродуктивных, предложений русского языка.

Литература

1. *Chafe W.* (ed.) (1980), The pear stories: Cognitive, cultural, and linguistic aspects of narrative production, Norwood: Ablex.
2. *Chafe W.* (1994), Discourse, consciousness, and time. The flow and displacement of conscious experience in speaking and writing, Chicago.
3. *Du Bois J.* (1980), The trace of identity in discourse, W. Chafe (Ed.) The Pear Stories, New Jersey: Ablex, pp. 1–7.
4. *Fedorova O. V.* (1994), A typology of referent introduction means [Tipologiya sredstv introduktsii referenta], Moscow.
5. *Kibrik A. A., Podlesskaya V. I.* (eds.) (2009). Corpus of spoken Russian “Night Dream Stories” [Korpus ustnoy russkoy rechi “Rasskazy o snovideniyakh”], Moscow: Jazyki slavyanskikh kul’tur.
6. *Mazur I., Chmiel A.* (2012), Towards common European audio description guidelines: Results of the Pear Tree Project, Perspectives: Studies in Translatology, Vol. 20(1), pp. 5–23.
7. *Tannen D.* (1980), A comparative analysis of oral narrative strategies: Athenian Greek and American English, W. Chafe (ed.) The Pear Stories, New Jersey: Ablex, pp. 51–87.

ПОХОЖИ ЛИ РИТОРИЧЕСКИЕ СТРУКТУРЫ ДОКУМЕНТА И МЕТА-ДОКУМЕНТА?

Галицкий Б. А. (bgalitsky@hotmail.com)

Кноуледж-трэйл, Сан Хосэ, Калифорния, США

Формулируется проблема классификации текста по принадлежности к документу или мета-документа (паттерны метаязыка и языка-объекта), а также предлагаются ее области применения. Применяется метод ядер на расширенных деревьях разбора, полученных в резльтате склейки деревьев для предложений на основе анафоры, риторических структур и коммуникативных действий. Мы оцениваем наш подход с помощью корпуса инженерных документов, а также в области литературы. Предложенный метод позволяет надежно различать тексты с паттернами на языке-объекте и на метазыке, опираясь в основном на соответствующие риторические структуры.

Ключевые слова: метод ядер на расширенных деревьях разбора, язык-объект и метазык

DOCUMENT VS. META-DOCUMENT: ARE THEIR RHETORIC STRUCTURES DIFFERENT?

Galitsky B. A. (bgalitsky@hotmail.com)

Knowledge-Trail Inc. San Jose CA USA

The problem of classifying text with respect to belonging to a document or a meta-document (metalanguage and language object patterns) is formulated and its application areas are proposed. An algorithm is proposed for document classification tasks where counts of words is insufficient do differentiate between such abstract classes of text as metalanguage and object-level. We extend the parse tree kernel method from the level of individual sentences towards the level of paragraphs, based on anaphora, rhetoric structure relations and communicative actions linking phrases in different sentences. Tree kernel learning is then applied to these extended trees to leverage of additional discourse-related information. We evaluate our approach in the domain of action-plan documents, as well as in literature domain, recognizing some portions of text in Kafka's novel "The Trial" as metalanguage patterns and differentiating them from the novel's description in the studies of Kafka by others.

Key words: rhetoric structure, metalanguage, tree kernel, semantic discourse

1. Introduction

Solving text classification problems, keywords and their topicality usually suffice. These features provide abundant information to determine a topic of a text or document, such as apple vs banana, or adventures vs relaxing travel. At the same time, there is a number of document classification domains where distinct classes have similar words. In this case, style, phrasings and other kinds of text structure information need to be leveraged. To perform text classification in such domains, one needs to employ discourse information such as anaphora, rhetoric structure, entity synonymy and ontology, if available [11].

In this study, an issue of classifying a text with respect to being metalanguage or language object is addressed. We are concern with differentiating between object-level documents, which inform us on how to do things, or how something has been done, and meta-documents, specifying how to write a document which explains how to do things, or how things have been done. Metalanguage is a symbolic system intended to express information, or analyze another language or symbolic system. In proof theory, metalanguage is a language in which proofs are dealt with. Conversely, object-level the logic itself. In logic, it is a language in which the truth of statements in another language is being discussed. Logic programs can be recognized as meta-programs or object-level programs easily [4]. We refer to meta-document as a document whose text extensively uses a metalanguage.

In a natural language document, metalanguage is used as a special expressive means to ascend to the desired level of abstraction. To automatically recognize metalanguage patterns in text one, needs some implicit signals at the syntactic level. Naturally, just using keyword statistics is insufficient to differentiate between texts in metalanguage and language-object.

A presence of verbs for speech acts and mental states (such as knowing) may help to identify metalanguage patterns, but is an unreliable criterion*: I know the location of the highest mountain* vs *I know what he thinks about the highest mountain in the world*. The latter sentence contains a meta-predicate *think (who, about-what)* with the second variable ranging over a set of (object-level) expressions for thoughts about the *highest mountain*. Relying on syntactic parse trees would provide us with specific expressions and phrasings connected with a metalanguage. However, it will still be insufficient for a thorough description of linguistic features inherent to a metalanguage. It is hard to identify such features without employing a discourse structure of a document. This discourse structure needs to include anaphora, rhetoric relations, and interaction scenarios by means of communicative language[7]. Furthermore, to systematically learn these discourse features associated with metalanguage, and differentiate them from the ones for language-object, one needs a unified approach to classify graph structures at the level of paragraphs [5, 6].

The design of such features for automated learning of syntactic and discourse structures for classification is still done manually today. To overcome this problem, tree kernel approach has been proposed [1]. Tree kernels constructed over syntactic parse trees, as well as discourse trees [10] is one of the solutions to conduct feature engineering. Convolution tree kernel [3, 12] defines a feature space consisting of all subtree

types of parse trees and counts the number of common subtrees to express the respective distance in the feature space. They have found a broad range of applications in NLP tasks such as syntactic parsing re-ranking, relation extraction [16], named entity recognition [1], pronoun resolution [13], question classification, and machine translation.

The kernel ability to generate large feature sets is useful to assure we have enough linguistic features to differentiate between the classes, to quickly model new and not well understood linguistic phenomena in learning machines. However, it is often possible to manually design features for linear kernels that produce high accuracy and fast computation time whereas the complexity of tree kernels may prevent their application in real scenarios. SVM [25] can work directly with kernels by replacing the dot product with a particular kernel function. This useful property of kernel methods, that implicitly calculates the dot product in a high-dimensional space over the original representations of objects such as sentences, has made kernel methods an effective solution to modeling structured linguistic objects.

An approach to build a kernel based on more than a single parse tree for search has been proposed [9, 10]. To perform classification based on additional discourse features, we form a single tree from a tree forest for a sequence of sentences in a paragraph of text. Currently, kernel methods tackle individual sentences. For example, in question answering, when a query is a single sentence and an answer is a single sentence, these methods work fairly well. However, in learning settings where texts include multiple sentences, we need to represent structures which include paragraph-level information such as discourse.

A number of NLP tasks such as classification require computing of semantic features over paragraphs of text containing multiple sentences. Doing it at the level of individual sentences and then summing up the score for sentences will not always work. In the complex classification tasks where classes are defined in an abstract way, the difference between them may lay at the paragraph level and not at the level of individual sentences. In the case where classes are defined not via topics but instead via writing style, discourse structure signals become essential. Moreover, some information about entities can be distributed across sentences, and classification approach needs to be independent of this distribution. We will demonstrate the contribution of paragraph-level approach vs the sentence level in our evaluation.

2. The domain of documents and meta-documents

Our first example of the use of meta-language is the following text shared by an upset customer, doing his best to have a bank to correct an error: *The customer representative acknowledged that the only thing he is authorized to do is to inform me that he is not authorized to do anything...* This is a good example for how people describe *thinking about thinking.* In this example, bank operations can be described in language-object, and bank employee's authorizations to perform these operations are actually described in metalanguage. Here a document on banking operations is an object-level document, and authorization rules document is a meta-document relative to the operations document. The claim of this work is that this classification can be performed based on text analysis only without any knowledge of banking industry.

We define an action-plan (object-level) document as a document which contains a thorough and well-structured description of how to build a particular system or work of art, from engineering to natural sciences to creative art. According to our definition, action-plan document follows the reproducibility criteria of a patent or research publication; however format might deviate significantly. One can read such document and being proficient in the knowledge domain, can build such a system or work of art.

Conversely, a meta-document is a document explaining how to write object-level, action-plan documents. They include manuals, standard action-plan documents should adhere to, tutorials on how to improve them, and others. We need to differentiate action-plan documents from the classes of documents which can be viewed as ones containing meta-language, whereas the genuine action-plan documents consists of the language-object patterns and should not include metalanguage ones. As to the examples of meta-documents, they include design requirements, project requirement document, operational requirements, design guidelines, design guides, tutorials, design templates (template for technical design document, research papers on system design, educational materials on system design, resume of a design professional, and others.

Naturally, action-plan documents are different from similar kinds of documents on the same topic in terms of style and phrasing. To extract these features, rhetoric relations are essential. Notice that meta-documents can contain object-level text, such as design examples. Object level documents (genuine action-plan docs) can contain some author reflections on the system design process (which are written in metalanguage). Hence the boundary between classes does not strictly separates metalanguage and language object. We use statistical language learning to optimize such boundary, having supplied it with a rich set of linguistic features up to the discourse structures. In the design document domain, we will differentiate between texts expressed mostly via meta-language and the ones mostly in language-object.

A combination of object-language and metalanguage patterns and description styles can also be found in literature. Describing the nature, a historical event, an encounter between people, an author uses a language object. Describing the thought, beliefs, desires and knowledge of characters about the nature, events and interactions between people, an author may use a metalanguage, if its entities/range over the expressions (phrases) of the language-object.

An outstanding example of the use of metalanguage in literature is Franz Kafka's novel "The Trial". According to our model, the whole plot is described in metalanguage, and object-level layer is not presented at all. This is unlike a typical work of literature, where both levels are employed and object-level prevail, such as fairy tales. In "The Trial" we find out that the main character Joseph is being prosecuted, his thoughts and feelings are described. Also, his meeting with various people related to the trial are presented, but they are not attached to the essence of what was happening . No information is available about a reason for the trial, the charge, and the circumstances of the deed (that would be a language-object level information). The novel is a pure example of the presence of meta-theory and absence of object-level theory, from the standpoint of logic. The reader is expected to form the object–level theory herself to avoid an ambiguity in the interpretation of this novel.

Use of "The Trial" text as a training dataset would assist in understanding the linguistic properties of metalanguage and language-object. For example, it is easy to differentiate between a mental and a physical word, just relying on keywords. However, to distinguish meta-language from language object in text, one need to consider different discourse structures, which we will automatically learn from text.

In the literature domain, we will attempt to draw a boundary between the pure metalanguage (works of literature with a special level of abstraction) and a mixed level text (a typical work of literature).

3. Learning discourse structure via tree kernels

It turns out that sentence-level tree kernels are insufficient for classification in our domains. Since important phrases can be distributed through different sentences, one needs a sentence boundary—independent way of extracting both syntactic and discourse features. Therefore we intend to combine/merge parse trees to make sure we cover all the phrase of interest. Let us analyze the following text with respect of belonging to a document or meta-document.

> *This document describes the design of back end processor. Its requirements are enumerated below.*

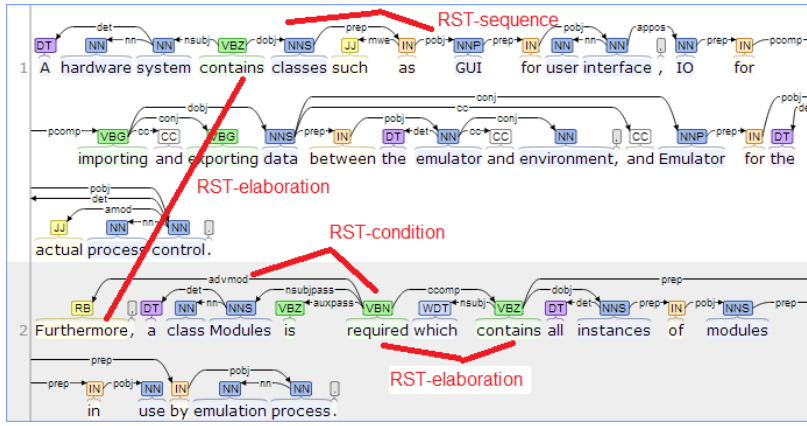
From the first sentence, it looks like an action-plan document. To process the second sentence, we need to disambiguate the preposition 'its'. As a result, we conclude from the second sentence that it is a requirements document, not an object-level action-plan one.

The structure of a document which can be potentially valuable for classification can be characterized by rhetoric relations that hold between the parts of a text. These relations, such as explanations or contrast, are important for text understanding in general since they contain information on how these parts of text are related to each other to form a coherent discourse. Naturally, we expect the structure of discourse for metalanguage text patterns to be different to that of language-object text patterns.

Rhetorical Structure Theory (RST, [15, 18]) is one of the most popular approaches to model extra-sentence as well as intra-sentence discourse. RST represents texts by labeled hierarchical structures, called Discourse Trees (DTs). The leaves of a DT correspond to contiguous Elementary Discourse Units (EDUs). Adjacent EDUs are connected by rhetorical relations (e.g., Elaboration, Contrast), forming larger discourse units (represented by internal nodes), which in turn are also subject to this relation linking. Discourse units linked by a rhetorical relation are further distinguished based on their relative importance in the text: nucleus being the central part, whereas satellite being the peripheral one. Discourse analysis in RST involves two subtasks: discourse segmentation is the task of identifying the EDUs, and discourse parsing is the task of linking the discourse units into a labeled tree. Discourse analysis explores how meanings can be built up in a communicative process, which varies between a text metalanguage and a text language-object. Each part of a text has a specific role in conveying the overall message of a given text.

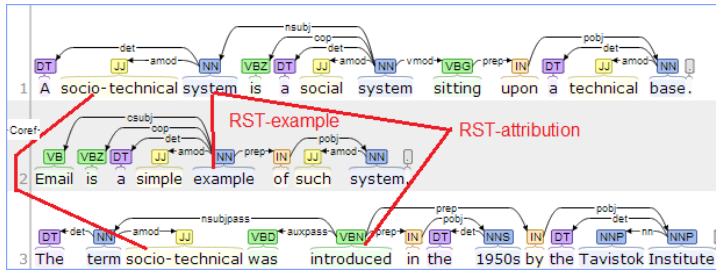
For our classification tasks, just an analysis of a text structure can suffice for proper classification. Given a positive sequence

A hardware system contains classes such as GUI for user interface, IO for importing and exporting data between the emulator and environment, and Emulator for the actual process control. Furthermore, a class Modules is required which contains all instances of modules in use by emulation process.



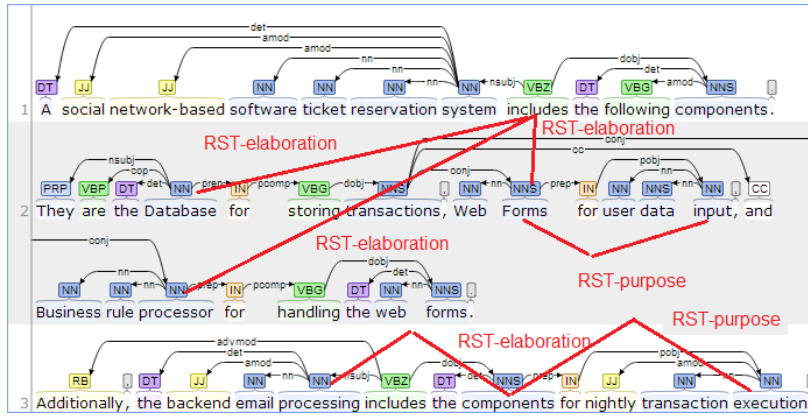
and a negative sequence

A socio-technical system is a social system sitting upon a technical base. Email is a simple example of such system. The term socio-technical was introduced in the 1950s by the Tavistok Institute.



We want to classify the paragraph

A social network-based software ticket reservation system includes the following components. They are the Database for storing transactions, Web Forms for user data input, and Business rule processor for handling the web forms. Additionally, the backend email processing includes the components for nightly transaction execution.



One can see that it follows the rhetoric structure of the top (positive) training set element, although it shares more common keywords with the bottom (negative) element. Hence we classify it as an action-plan document, being an object-level text, since it describes the system rather than introduces a terms (as the negative element does).

4. Anaphora and rhetoric relations for classification tasks

We introduce a classification problem where keyword and even phrase-based features are insufficient. This is due to the variability of ways information can be communicated in multiple sentences, and variations in possible discourse structures of text which needs to be taken into account.

We consider an example of text classification problem, where short portions of text belong to two classes:

- Tax liability of a landlord renting office to a business.
- Tax liability of a business owner renting an office from landlord.

I rent an office space. This office is for my business. I can deduct office rental expense from my business profit to calculate net income.

To run my business, I have to rent an office. The net business profit is calculated as follows. Rental expense needs to be subtracted from revenue.

To store goods for my retail business I rent some space. When I calculate the net income, I take revenue and subtract business expenses such as office rent.

I rent out a first floor unit of my house to a travel business. I need to add the rental income to my profit. However, when I repair my house, I can deduct the repair expense from my rental income.

I receive rental income from my office. I have to claim it as a profit in my tax forms. I need to add my rental income to my profits, but subtract rental expenses such as repair from it.

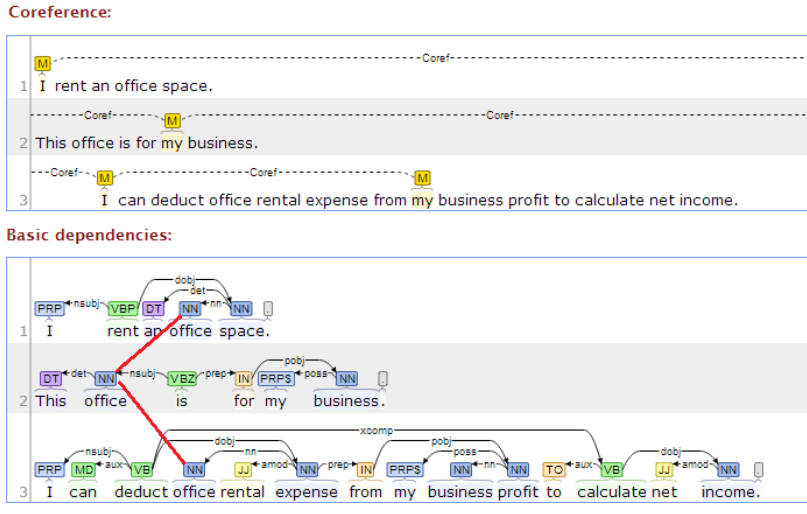
I advertised my property as a business rental. Advertisement and repair expenses can be subtracted from the rental income. Remaining rental income needs to be added to my profit and be reported as taxable profit.

Note that keyword-based analysis does not help to separate the first three paragraph and the second three paragraphs. They all share the same keywords *rental/office/income/profit/add/subtract.* Phrase-based analysis does not help, since both sets of paragraphs share similar phrases.

Secondly, pair-wise sentence comparison does not solve the problem either. Anaphora resolution is helpful but insufficient. All these sentences include 'I' and its mention, but other links between words or phrases in different sentences need to be used.

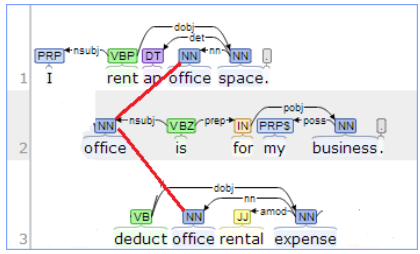
Rhetoric structures need to come into play to provide additional links between sentences. The structure to distinguish between *renting for yourself and deducting from total income* and *renting to someone and adding to income* embraces multiple sentences. The second clause about *adding/subtracting incomes* is linked by means of the rhetoric relation of *elaboration* with the first clause for *landlord/tenant*. This rhetoric relation may link discourse units within a sentence, between consecutive sentences and even between first and third sentence in a paragraph. Other rhetoric relations can play similar role for forming essential links for text classification.

Which representations for these paragraphs of text would produce such common sub-structure between the structures of these paragraphs? We believe that extended trees, which include the first, second, and third sentence for each paragraph together can serve as a structure to differentiate the two above classes. The dependency parse trees for the first text in our set and its coreferences are shown below:



There are multiple ways the nodes from parse trees of different sentences can be connected: we choose the rhetoric relation of elaboration which links the same entity office and helps us to form the structure *rent-office-space—for-my-business—deduct-rental-expense* which is the base for our classification.

We show the resultant extended tree with the root 'I' from the first sentence.



It includes the whole first sentence, a verb phrase from the second sentence and a verb phrase from the third sentence according to rhetoric relation of elaboration. Notice that this extended tree can be intuitively viewed as representing the 'main idea' of this text compared to other texts in our set. All extended trees need to be formed for a text and then compared with that of the other texts, since we don't know in advance which extended tree is essential. From the standpoint of tree kernel learning, extended trees are learned the same way as regular parse trees.

5. Building extended trees and learning them

For every inter-sentence arc which connects two parse trees, we derive the extension of these trees, extending branches according to the arc (Fig. 1).

In this approach, for a given parse tree, we will obtain a set of its extension, so the elements of kernel will be computed for many extensions, instead of just a single tree. The problem here is that we need to find common sub-trees for a much higher number of trees than the number of sentences in text, however by subsumption (sub-tree relation) the number of common sub-trees will be substantially reduced.

If we have two parse trees P_1 and P_2 for two sentences in a paragraph, and a relation R_{12}: $P_{1i} \rightarrow P_{2j}$ between the nodes P_{1i} and P_{2j}, we form the pair of extended trees $P_1 * P_2$:

$$\ldots, P_{1i\text{-}2}, P_{1i\text{-}1}, P_{1i}, P_{2j}, P_{2j+1}, P_{2j+2}, \ldots$$

$$\ldots, P_{2j\text{-}2}, P_{2j\text{-}1}, P_{2j}, P_{1i}, P_{1i+1}, P_{2i+2}, \ldots,$$

which would form the feature set for tree kernel learning in addition to the original trees P_1 and P_2.

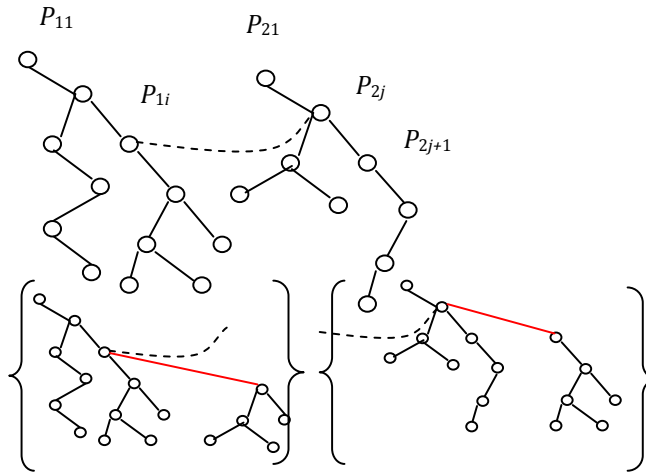


Fig. 1: An arc which connects two parse trees for two sentences in a text (on the top) and the derived set of extended trees (on the bottom)

The algorithm for building an extended tree for a set of parse trees T is presented below:

Input:
1) Set of parse trees T.
2) Set of relations R, which includes relations R_{ijk} between the nodes of T_i and T_j: $T_i \in T$, $T_j \in T$, $R_{ijk} \in R$. We use index k to range over multiple relations between the nodes of parse tree for a pair of sentences.

Output: the exhaustive set of extended trees E.

Set $E = \varnothing$;
For each tree i=1:|T|
  For each relation R_{ijk}, k = 1:|R|
    Obtain T_j
    Form the pair of extended trees $T_i * T_j$;
    Verify that each of the extended trees do not have a super-tree in E
    If verified, add to E;
Return E.

Notice that the resultant trees are not the proper parse trees for a sentence, but nevertheless form an adequate feature space for tree kernel learning.

Kernel methods are a large class of learning algorithms based on inner product vector spaces. Support vector machines (SVMs) are mostly well-known algorithms. The main idea behind SVMs is to learn a hyperplane,

$$H(\vec{x}) = \vec{w} \cdot \vec{x} + b = 0$$

where $\vec{x}$ is the representation of a classifying object o as a feature vector, while $\vec{w} \in \Re^n$ (indicating that $\vec{w}$ belongs to a vector space of n dimensions built on real numbers) and $b \in \Re$ are parameters learned from training examples by applying the Structural Risk Minimization principle (Cortez & Vapnik 1995).

Convolution kernels as a measure of similarity between trees compute the common sub-trees between two trees T_1 and T_2. Convolution kernel does not have to compute the whole space of tree fragments. Let the set $\tau = \{t_1, t_2, \ldots, t_{|\tau|}\}$ be the set of sub-trees of an extended parse tree, and $\chi_i(n)$ be an indicator function which is equal to 1 if the subtree t_i is rooted at a node n, and is equal to 0 otherwise. A tree kernel function over trees T_1 and T_2 is

$$TK(T_1, T_2) = \sum_{n_1 \in N_{T_1}} \sum_{n_1 \in N_{T_2}} \Delta(n_1, n_2),$$

where N_{T_1} and N_{T_2} are the sets of T_1's and T_2's nodes, respectively and

$$\Delta(n_1, n_2) = \sum_{i=1}^{|\tau|} \chi_i(n_1)\chi_i(n_2)$$

It calculates the number of common fragments with the roots in and nodes.

There are following processing steps used in our classifier. Each paragraph of a document is subject to sentence splitting, part-of-speech tagging, dependency parsing and chunking. We also rely on additional tags to extend SVM feature space, finding similarities between trees. These additional tags include noun entities from Stanford NLP such as organization and title, and verb types from VerbNet. We then produce a graph-based representation for a document, applying anaphora and our own RST parser for inter-sentence relations.

To obtain the inter-sentence links, we employ coreferences from Stanford NLP [17, 20]. Rhetoric relation extractor is based on our rule-based approach to finding relations between EDUs [10]. We combine manual rules with automatically learned rules derived from the available discourse corpus by means of syntactic generalization. For each inter-sentence arc between two parse trees, we form a pair of extended trees from the source and destination parse trees for this arc [6]. Finally, we form a training dataset of extended trees and pass it on to SVM Parse tree kernel learner [14].

6. Evaluation

For the action-plan document domain, we formed a set of 940 action-plan documents from the web. We also compiled the set of meta- documents on similar engineering topics, mostly containing the same keywords. For the literature domain, we collected 160 paragraphs as meta-documents from Kafka's novel "The Trial" as well as his other novels so that these paragraphs are read as metalanguage patterns. As a set of object-level documents we manually selected 200 paragraphs of text

in the same domain (scholarly articles about "The Trial"). We split the data into 3 subsets for training/evaluation portions and cross-validation [19].

Table 1 shows evaluation results for the both above domains. Each row shows the results of the baseline classification methods, such as Keyword statistics (TF*IDF, [21, 22], Nearest-Neighbor classification and Naïve Bayes [23, 24].

Baseline approaches show rather low performance. The one of the tree kernel based methods improves as the sources of linguistic properties are expanded. For both domains, there is an improvement by a few percent due to the rhetoric relations compared with the baseline tree kernel SVM which employs parse trees only. For the literature documents, the role of anaphora is lower than for technical ones.

Table 1: Classifying text into metalanguage and language-object

	Actin-plan document, %			Literature doc		
Method	**Precision**	**Recall**	**F-measure**	**Precision**	**Recall**	**F-measure**
Nearest neighbor classifier–TF*IDF based	53.9	62.0	57.67	48.5	54.3	51.24
Naive Bayesian classifier (WEKA)	55.3	59.7	57.42	50.6	51.0	50.80
Tree kernel—regular parse trees	71.4	76.9	74.05	63.3	68.7	65.89
Tree kernel SVM—extended trees for anaphora	77.8	81.4	79.56	69.3	65.6	67.40
Tree kernel SVM—extended trees for RST	80.1	80.5	80.30	69.8	74.5	72.07
Tree kernel SVM—extended trees for both anaphora & RST	83.3	83.6	83.45	71.5	73.1	72.29

7. Related work and conclusions

We have previously studied enriching a set of linguistic information such as syntactic relations between words helps in search and other relevance tasks [6,8]. To leverage semantic discourse information and especially rhetoric relations, we introduced parse thicket representation of documents and defined generalization operation on parse thickets [9]. We also proposed how the feature space of tree kernel learning can be expanded to accommodate for semantic discourse features [10].

In this study we addressed the issue of how semantic discourse features assist with solving such abstract classification problem as differentiating between natural language-object and natural meta-language. We demonstrated that the problem of such level of abstraction can nevertheless be dealt with statistical learning allowing automated feature engineering. Evaluation domains are selected so that the only

differences between classes are in phrasing and discourse structures (not in keywords). We also demonstrated that both of these structures are learnable.

We draw the comparison with two following sets of linguistic features:

1) The baseline set, parse trees for individual sentences,
2) Parse trees and discourse information,

and showed that the enhanced set indeed improves the classification performance for the same learning framework. One can see that the baseline text classification approaches does not perform well in the classification domain as abstract and complicated as recognizing metalanguage.

A number of studies explored various forms of meta-language and meta-reasoning, however to the best of our knowledge a system which automatically recognizes natural metalanguage has not being built. [2] proposed a fairly general approach to meta-reasoning as providing a basis for selecting and justifying computational actions. Addressing the problem of resource-bounded rationality, the authors provide a means for analyzing and generating optimal computational strategies. Because reasoning about a computation without doing it necessarily involves uncertainty as to its outcome, the authors select probability and decision theory as their main tools.

A system needs to implement metalanguage to impress peers of being humanlike and intelligent, being capable of thinking about one's own thinking. Traditionally within cognitive science and artificial intelligence, thinking or reasoning has been cast as a decision cycle within an action-perception loop [27]. An intelligent agent perceives some external world stimuli and responds to achieve its goals by selecting some action from its available set. The result of these actions at the ground level is subsequently perceived at the object level and the cycle continues. Meta-reasoning is the process of reasoning about this cycle. It consists of both the meta-level control of computational activities and the introspective monitoring of reasoning. In this study we focused on linguistic issues of texts which describe such cognitive architecture. We found an inter-connection between a cognitive architecture and a discourse structure used to express it in text. Relying on this inter-connection, one can automatically classify texts with respect to the cognitive level they describe.

In our previous studies we considered the following sources of relations between words in sentences: coreferences, taxonomic relations such as sub-entity, partial case, predicates for subject etc., rhetoric structure relations, and speech acts [7]. We demonstrated that a number of NLP tasks including search relevance can be improved if search results are subject to confirmation by parse thicket generalization, when answers occur in multiple sentences. In this study we employed coreferences and rhetoric relation only to identify correlation with the occurrence of metalanguage in text. Although phrase-level analysis allows extraction of weak correlation with metalanguage in text, ascend to discourse structures makes detection of metalanguage more reliable. In our evaluation setting, using discourse improved the classification F-measure by 5.5–8.6% depending on a classification sub-domain.

There is a strong dis-attachment between modern text learning approaches and text discourse theories. Usually, learning of linguistic structures in NLP tasks is limited to keyword forms and frequencies. On the other hand, most theories of semantic discourse are not computational in nature. In this work we attempted to achieve the

best of both worlds: learn complete parse tree information augmented with an adjustment of discourse theory allowing computational treatment.

In this paper, we used extended parse trees instead of regular ones, leveraging available discourse information, for text classification. This work describes one of the first applications of tree kernel to industrial scale NLP tasks. The advantage of this approach is that the manual thorough analysis of text can be avoided for complex text classification tasks where the classes are as high-level as documents vs meta-documents. The reason of the satisfactory performance of the proposed classification method is a robustness of statistical learning algorithms to noisy and inconsistent features extracted from documents.

The experimental environment, extended tree learning functionality and the evaluation framework is available at http://code.google.com/p/relevance-based-on-parse-trees.

References

1. *Cumby, C. and Roth D.* (2003) On Kernel Methods for Relational Learning. ICML, pp. 107–14.
2. *Russell, S., Wefald, E., .Karnaugh, M., Karp, R., McAllester, D., Subramanian, D., Wellman, M.* (1991) Principles of Metareasoning, Artificial Intelligence, pp. 400–411, Morgan Kaufmann.
3. *Collins, M., and Duffy, N.* (2002) Convolution kernels for natural language. In Proceedings of NIPS, 625–32.
4. *Galitsky, B.* (2003) Natural Language Question Answering System: Technique of Semantic Headers. Advanced Knowledge International, Adelaide, Australia.
5. *Galitsky, B., de la Rosa J. L., Dobrocsi, G.* (2012) Inferring the semantic properties of sentences by mining syntactic parse trees. Data & Knowledge Engineering. Volume 81–82, November 21–45.
6. *Galitsky, B., Usikov, D., and Kuznetsov S. O.* (2013) Parse Thicket Representations for Answering Multi-sentence questions. 20th International Conference on Conceptual Structures.
7. *Galitsky, B., Kuznetsov S.* (2008) Learning communicative actions of conflicting human agents. *J. Exp. Theor. Artif. Intell.* 20(4): 277–317 .
8. *Galitsky, B.* (2012) Machine Learning of Syntactic Parse Trees for Search and Classification of Text. Engineering Applications of Artificial Intelligence. 26 (3), 1072–91
9. *Galitsky, B.* (2014) Learning parse structure of paragraphs and its applications in search. Engineering Applications of Artificial Intelligence. 32, 160–84.
10. *Galitsky B., Ilvovsky, D., Kuznetsov, S. O.* (2015) Text Integrity Assessment: Sentiment Profile vs Rhetoric Structure. CICLing-2015, Cairo, Egypt.
11. *Wu, J., Xuan Z. and Pan, D.* (2011) Enhancing text representation for classification tasks with semantic graph structures, International Journal of Innovative Computing, Information and Control (ICIC), Volume 7, Number 5(B).
12. *Haussler, D.* (1999) Convolution kernels on discrete structures. UCSB Technical report.

13. *Kong, F. and Zhou G.* (2011) Improve Tree Kernel-Based Event Pronoun Resolution with Competitive Information. Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence, 3 1814–19.
14. *Moschitti, A.* (2006) Efficient Convolution Kernels for Dependency and Constituent Syntactic Trees. 2006. In Proceedings of the 17th European Conference on Machine Learning, Berlin, Germany,
15. *Mann, W., Matthiessen C. and Thompson S.* (1992) Rhetorical Structure Theory and Text Analysis. Discourse Description: Diverse linguistic analyses of a fund-raising text. ed. by Mann W and Thompson S.; Amsterdam, John Benjamins. pp. 39–78.
16. *Zhang, M.; Che, W.; Zhou, G.; Aw, A.; Tan, C.; Liu, T.; and Li, S.* (2008) Semantic role labeling using a grammar-driven convolution tree kernel. IEEE transactions on audio, speech, and language processing. 16(7):1315–29.
17. *Lee, H., Chang, A., Peirsman, Y., Chambers, N., Surdeanu, M. and Jurafsky, D.* (2013) Deterministic coreference resolution based on entity-centric, precision-ranked rules. Computational Linguistics 39(4), 885–916.
18. *Marcu, D.* (1997) From Discourse Structures to Text Summaries, in I. Mani and M. Maybury (eds) Proceedings of ACL Workshop on Intelligent Scalable Text Summarization, pp. 82–8, Madrid, Spain.
19. *Kohavi, R.* (1995) A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. International Joint Conference on Artificial Intelligence. 1137–43.
20. *Recasens, M., de Marneffe M-C, and Potts, C.* (2013) The Life and Death of Discourse Entities: Identifying Singleton Mentions. In Proceedings of NAACL.
21. *Croft, B., Metzler, D., Strohman, T.* (2009) Search Engines — Information Retrieval in Practice. Pearson Education. North America.
22. *Salton, G. and Buckley, C.* (1988) Term-weighting approaches in automatic text retrieval. Information Processing & Management 24(5): 513–23.
23. *Moore, J. S. and Boyer R. S.* (1991) MJRTY — A Fast Majority Vote Algorithm, In R. S. Boyer (ed.), Automated Reasoning: Essays in Honor of Woody Bledsoe, Automated Reasoning Series, Kluwer Academic Publishers, Dordrecht, The Netherlands, pp. 105–17.
24. *John G. H. and Langley, P.* (1995) Estimating Continuous Distributions in Bayesian Classifiers. In Eleventh Conference on Uncertainty in Artificial Intelligence, San Mateo, 338–45.
25. *Vapnik, V.* (1995) The Nature of Statistical Learning Theory, Springer-Verlag.
26. *Michael T. Cox and Anita Raja.* (2007) Metareasoning: A manifesto.

ЛИНГВИСТИЧЕСКИЙ АНАЛИЗ В СИСТЕМАХ ИДЕНТИФИКАЦИИ ДИКТОРА: ИНТЕГРАТИВНЫЙ КОМПЛЕКСНЫЙ ПОДХОД НА БАЗЕ ЭКСПЕРТОЛОГИИ

Галяшина Е. И. (galyashina@gmail.com)

Московский государственный университет
им. О. Е. Кутафина, Москва, Россия

В статье предлагается концепция экспертной технологии идентификации диктора посредством интегрирования и комплексирования перцептивных, акустико-фонетических и собственно лингвистических методов анализа звучащей речи, определяющих общую и специальную компетенцию речеведа. Автор анализирует ключевые понятия и термины процедуры индивидуально-конкретного отождествления диктора в аспекте современной экспертологии. Показано, что в условиях, когда сравнению подвергаются оцифрованные речевые сигналы, полученные по неизвестному алгоритму преобразования и при невозможности исключить их фальсификацию, для достоверного и надежного решения идентификационной задачи возрастает роль профессиональных лингвистических компетенций, которые должны обладать достаточной интегративной широтой междисциплинарного знания и комплексностью познания феномена индивидуальности речевого следа.

Ключевые слова: лингвистический анализ, идентификация диктора, экспертные системы, интегрированные знания, экспертология

LINGUISTIC ANALYSIS IN THE SPEAKER IDENTIFICATION SYSTEMS: INTEGRATED COMPLEX EXAMINATION APPROACH BASED ON FORENSIC SCIENCE TECHNOLOGY

Galyashina E. I. (galyashina@gmail.com)

Kutafin Moscow State Law University, Moscow, Russia

The article proposes the concept of the integrated expert techniques for speaker identification on the domain of complex acoustic-phonetic and linguistic methods of oral speech analysis, defining general and special forensic expert competences. The result of forensic speaker identification used as evidence must exhibit a high level of reliability. The author examines the key concepts and terms of the procedure of individual-specific speaker identification in the aspect of modern expertology (forensic science). The paper states the need to take into account that the role of professional linguistic competences increases in conditions when digitized speech signals are compared, algorithm of coding is indefinite and falsification of utterances is not excluded. To solve this problem the author proposes a multistage approach consisting of a parallel application of instrument and technical methods together with aural-perceptual, waveform and sonogram investigation and sophisticated linguistic analysis. The main attention is paid to the linguistic component of the complex integrated approach based on the phonetic and semantic analyses. It is stated that individualized speech unit is formed by a system of miscellaneous formal and semantic relations of structural speech components in linguistic contents. The proposed method of integration of the multilevel speech modules was implemented in forensic linguistic methodology of speaker identification technique. This made it possible to considerably increase the reliability of the expert's decision and provided an opportunity to use it as a component of the multistage system for speech utterances authentication.

Key words: linguistic analysis, speaker identification, expert systems, integrated expert techniques, forensics

Системный анализ лингвистических методов, применяемых в современных автоматизированных системах идентификации диктора, имеющих экспертную направленность, позволил установить сложную динамическую природу индивидуальности речевого следа, выступающего в качестве объекта экспертизы. Для его полного и всестороннего исследования требуются интегрированные знания эксперта, которые должны обладать достаточной широтой и вместе с тем глубиной познания исследуемого речевого объекта. При их использовании в практической деятельности формируются дополнительные профессиональные речеведческие компетенции, представляющие собой комплекс практических навыков и умений, получаемых опытным путем, закрепление которых должно входить в систему подготовки по соответствующей экспертной специализации. Компетенция, определяемая экспертной технологией отождествления диктора, это не механическое соединение разных областей научных знаний о речи, а комплексное интегративное образование, позволяющее сформировать требуемые профессиональные компетенции на базе современной экспертологии. Применительно к лингвистическому анализу это, в первую очередь, относится к компетентной оценке индивидуально-определенной специфики фонетической и смысловой организации высказываний в речевом потоке.

В настоящее время существует ряд методических подходов, базирующихся на применении автоматизированных экспертных систем текстозависимой и текстонезависимой идентификации диктора. Вместе с тем их возможности ограничены при исследовании речевого сигнала, продуцированного в условиях компьютерно-опосредованной речевой коммуникации, при образовании речевого

следа путем цифровой передачи сигнала по каналам мобильной связи, а также воздействии на структуру звукового сигнала при его цифровой обработке, сжатии и иных способах модификации. Существующие автоматические системы и правила принятия идентификационного решения не позволяют достигать результатов, достоверность, надежность и воспроизводимость которых отвечает требованиям обеспечения доказательств в практике российского судопроизводства[1]. Доля неавтоматизированного труда эксперта остается значительной, что снижает степень объективности вывода. Технология производства экспертизы предусматривает разрозненное использование методов акустического и лингвистического анализа, на практике реализуемого разными субъектами, обладающими различными компетенциями, независимо друг от друга. Общее решение о принадлежности речевого сигнала на проверяемой фонограмме конкретному диктору принимается двумя экспертами (инженером и лингвистом) с разным объемом специальных знаний путем суммирования результатов параметрического сравнения инженером противопоставленных сигналограмм и социолингвистического портрета, сформированного на уровне слуховой перцепции лингвиста. Лингвистические признаки структурно представлены перечнями описаний, не детерминированными по их специфичности. Не учитывается объективно-системный характер речи как бинарного процесса, опирающегося на индивидуальные психофизиологические процессы в организме человека и оформленного потоком языковой последовательности звуков с семантическим содержанием, дополняемым всевозможными интонационно-эмоциональными оттенками.

Можно констатировать, что единая общепринятая научно-обоснованная методика отождествления диктора пока отсутствует, а в экспертных организациях России изолированно применяются методы акустического и аудитивно-лингвистического анализа, механистически разрывающие неделимое знаковое единство речи как продукта языковой деятельности. Это приводит к том, что на одних и тех же объектах сравнения одним методом могут быть получены результаты, невоспроизводимые применением иного метода, либо эксперты могут формулировать прямо противоположные выводы, устанавливая тождество одним методом и отсутствие такого — другим. В экспертных заключениях не приводятся вероятностно-статистические данные о надежности принятия идентификационного решения (вероятности совершения ошибки первого, второго рода — «пропуска цели» или «ложного захвата цели») по отдельным видам исследований. Все применяемые методические материалы носят рекомендательный характер, не содержат строгих предписаний и формальных правил, оставляя эксперту широкое поле для субъективного усмотрения как по выбору количества и комбинаторике измеряемых параметров речевого сигнала, набору вычисляемых акустических признаков, списку интерпретируемых лингвистических признаков и их оценке, и в этом смысле собственно методиками не являются.

В экспертологии методика, понимаемая как «система категорических или альтернативных научно-обоснованных предписаний по выбору и применению

---

[1] См., например, Обзор методов идентификации на электронном ресурсе http://www.forenex.ru — дата последнего посещения 08.01.2015.

в определенной последовательности и в определенных существующих или создаваемых условиях методов, приемов и средств (приспособлений, приборов и аппаратуры) для решения экспертной задачи»[2], занимает в настоящее время одно из центральных мест. Методика может содержать несколько вариантов решения одной экспертной задачи с учетом специфики ее условий или особенностей представленных объектов, однако, каждый этап принятия экспертом конкретного решения в случае выбора альтернативных вариантов должен быть изложен в заключении эксперта и мотивирован.

Феномен экспертной методики как научно-обоснованный алгоритм решения конкретной практически ориентированной задачи, позволяющий формировать критерии для оценки достоверности заключения эксперта как доказательства, привлекает в условиях современных реалий судопроизводства особое внимание. Это тем более важно, что получаемые по имеющимся методическим подходам идентификации диктора результаты нередко невоспроизводимы и носят неоднозначный, противоречивый характер. Видимо, поэтому в своих заключениях эксперты часто не приводят ссылок на конкретную экспертную методику либо, подменяют понятие экспертной методики понятием «методические материалы», «методические рекомендации», смешивают общенаучные и частнонаучные методы, описывают содержание и ход проведенного исследования с перечислением типовых, а не конкретно использованных технических средств, не указывают логические основания вывода и правила принятия ими того или иного решения[3].

На примере, практики экспертного отождествления лица по фонограммам речи, этот тезис становится особенно актуальным. Обзор существующих методов и инструментов анализа речевого сигнала показывает, что для решения задачи идентификации диктора в отечественной экспертизе присутствуют разные подходы.

Первый реализован на основе автоматизированной системы идентификации дикторов «Диалект»[4], изложенный в соответствующих пособиях для экспертов[5]. В структуру функциональной схемы проведения идентификационных исследований на системе «Диалект» входят измерение спектрально-временных параметров гласных и согласных звуков, вычисление акустических признаков речи неизвестного лица на спорной фонограмме и образцах речи

---

[2] Россинская Е. Р., Галяшина Е. И., Зинин А. М. Теория судебной экспертизы. — Учебник. — М.: Норма, 2013, с. 130.

[3] Подробнее см. Россинская Е. Р., Галяшина Е. И. Настольная книга судьи: судебная экспертиза. — Москва: Проспект, 2014, с. 303–340.

[4] В настоящее время данный метод реализован в программе «Диалект», ее коммерческой версии «Фонэкси», а также в системе PhonoBase. Способ идентификации личности по фонограммам произвольной устной речи в части инструментального анализа был запатентован (патент № 2107950 от 08.08.1996, G10L 5/06: Байчаров Н. В., Карлин И. П., Кураченкова Н. Б., Линьков А. Н., Попов Н. Ф., Савельев Ю. И., Тимофеев И. Н., Фесенко А. В.).

[5] Идентификация лиц по устной речи на русском языке. Методика «Диалект». Пособие для экспертов. Издание 2-е, переработанное и дополненное / Н. Б. Кураченкова, Н. В. Байчаров, М. А. Ермакова. Под редакцией В. М. Богданова. — М., 2007.

подозреваемого лица, добавление в имеющуюся статистическую базу акустических признаков речи разных лиц информации о векторах признаков речи данного подозреваемого лица, переобучение системы, определение «веса» (информативности) совпадающих признаков и порогов допустимой вариативности акустических признаков. Система обладает возможностями адаптации процедуры принятия решения путем произвольного исключения из сравнения совпадающих или различающихся признаков, а также изменения порога принятия идентификационного решения. Однако, по мнению некоторых специалистов: «в целом данный метод соответствует уровню развития речевых технологий конца прошлого века и не учитывает многонационального характера современной преступности, объективных изменений в русском языке, появления и развития цифровых средств речевой связи»[6]. Отметим, что лингвистические признаки представлены в комплексе «Диалект» списком воспринимаемых на слух характеристик речевого потока, слова, фразы и звука, примеры звучания которых представлены в базе эталонов в системе. При этом эксперт выбирает из предлагаемых системой списка признаков те эталоны, которые в большей степени соответствуют его слуховому восприятию.

Второй подход основан на применении в качестве инструментария для измерения и вычисления речевых параметров различных звуковых редакторов, разработанных отечественными и зарубежными специалистами[7], которые требуют от пользователя высокой квалификации, поскольку методы обладают невысокой степенью автоматизации, в том числе блоками автоматического сравнения получаемых результатов с принятием идентификационного решения на уровне экспертной оценки. Лингвистический анализ при втором подходе сводится к слуховой перцепции и фонетическому описанию сегментных и суперсегментных особенностей произносительных навыков. Решение принимается экспертом методом субъективной оценки выявленных совпадений и различий.

В этом ряду особо следует отметить новую версию звукового редактора SIS II v2.0, в котором реализованы модули попарного и обобщенного идентификационного решения. В этом блоке пользователь имеет возможность по своему усмотрению менять границы применимости каждого из отдельных методов исследования, изменяя таким способом его весовой коэффициент в общем решении Лингвистический анализ представлен перечислением просодических особенностей с элементами субъективного психолого-социо-речевого портрета диктора[8].

---

[6] См., например, Обзор методов идентификации [электронный ресурс]: http://www.forenex.ru.

[7] Например, OTExpert (ООО «ОТ-Контакт», Москва), Justiphone (ООО «Целевые технологии», Москва), SIS II v2.0 (ООО «Центр речевых технологий», Санкт Петербург), а также Praat (Амстердамский университет), SFS (Speech Filing System) и др. Подробный список представлен в монографии Галяшина Е. И. Основы судебного речеведения. — М: Стэнси, 2003, с. 225–229.

[8] Булгакова Е. В., Краснова Е. В. Экспертные системы и методы идентификации диктора // ИЗВ. ВУЗОВ. ПРИБОРОСТРОЕНИЕ. 2014. Т. 57, № 2, с. 58–63.

Обзор современных методических подходов, используемых в экспертной практике показал:

- отсутствие терминологического единоообразия в применяемых подходах к решению задачи отождествления диктора с достаточной точностью и надежностью, приемлемой для целей доказывания в судопроизводстве;
- множественность подходов и отсутствие единой, унифицированной методики;
- изолированное использование методов акустического и слухового фонетического анализа независимо друг от друга с суммированием полученных результатов;
- применение методов акустического анализа, которые базируются на презумпции наличия у эксперта информации об источнике происхождения представленного на фонограмме следа речевого сигнала и путей (способов) его модификации от источника (диктора) до регистрации на носителе;
- лингвистический анализ представлен многомерным недетерминированным признаковым пространством;
- применение аудитивного (лингвистического) анализа базируется на презумпции, что реплицированная диалогическая коммуникация, атрибутирована по принадлежности ее участникам;
- информационная значимость алгоритмически определяется для совпадающих признаков, вес различающихся признаков устанавливается экспертом субъективно;
- наличие в автоматизированных системах адаптивных процедур, позволяющих вмешиваться в процедуру принятия решения;
- общее решение принимается эвристически, носит во многом еще субъективный характер и в значительной мере зависит от опыта и багажа знаний эксперта.

Проблема осложняется тем, что надежность идентификационного решения во многом зависит от качества речевого сигнала, записанного на материальный носитель, тогда как речь человека при прохождении через каналы цифровой мобильной связи неизбежно искажается[9]. Идентификация диктора в сигнале, подвергшемся модификации, кодированию, сжатию и иным видам цифрового (а иногда и аналогового) преобразования значительно затруднена[10]. При передаче речи по каналам мобильной связи экспертному исследованию подвергается

---

[9] Галяшина Е. И., Галяшин В. Н. Цифровые фонограммы как судебное доказательство, Воронежские криминалистические чтения, — Воронеж: Изд-во Воронежского гос. университета, № 8, 2007, с. 71.

[10] На практике нередко возникает задача установления личности анонимного свидетеля, голос которого намеренно искажается по неизвестному алгоритму с целью затруднения или исключения возможности его опознания на слух или идентификации по акустико-фонетическим параметрам. Доступным для исследования при этом остается уровень языковой организации речи, отражающий индивидуальную характеристику манеры речи говорящего, обусловленной спецификой образа мышления, понимания темы, полноты аргументации, связности, цельности, информативности речевого сообщения.

синтезированный речевой сигнал, из которого алгоритмом кодирования исключены (искажены) существенные идентифицирующие диктора признаки.

Методические требования к фонограммам, в которых сигнал подвергался дискретизации, цифровому преобразованию, сжатию по неизвестному экспертам алгоритму, ставят их в разряд объектов, в которых установить экспертными методами достоверность отображения первичной информации о голосовом источнике невозможно. Без решения вопросов о степени влияния каналов связи, алгоритмов цифровой обработки на каждую группу признаков, участвующих в процедуре идентификации диктора экспертное исследование речевого следа не может быть надежным и достоверным.

При использовании цифровых каналов передачи сигнала на носителе фиксируется не сам исходный объект (например, звуковая волна) или его полное отражение (например, электрический сигнал от микрофона), а его абстрактная (математическая) модель. По каналу связи передается не нативная речь, а некий набор кодов и символов по которым на оконечном аппарате восстанавливается (синтезируется) сигнал, похожий на исходный, но который не является исходным сигналом. Причем эта абстрактная модель характеризуется двумя основными элементами — видом используемой математической модели и ее параметрами. Когда решают вопрос о выборе абстрактной модели для описания реальной акустической волны, то во главу вопроса ставят сложность устройства записи, удобство передачи полученной цифровой записи по системам связи и ее хранения на материальных носителях, но никак не требования максимально полного сохранения ее индивидуальных признаков значимых для решения задач отождествления диктора.

Речевой сигнал считается пригодным для идентификации диктора при обработке его алгоритмами сжатия с определенной нижней границей скорости цифрового потока, а именно:

- 32 кбит/с для ADPCM;
- 9,6 кбит/с для линейных предсказателей.

Речевой материал может быть признан условно пригодным для идентификации диктора, когда эти значения могут составлять:

- 16 кбит/с для ADPCM;
- 8 кбит/с для линейных предсказателей.

Во остальных случаях кодированный речевой сигнал должен быть признан, не пригодным для экспертной идентификации диктора как физическими методами, так и лингвистическими, поскольку фонетическая структура речевого отрезка, прошедшего такую обработку, существенно отличается от оригинала[11].

---

[11] Желудков Р. Н., Тимко Е. В., Усков К. Ю. О влиянии сжатия речи на допустимость речевой фонограммы в уголовное судопроизводство. Материалы 2-ой Всероссийской конференции «Теория и практика речевых исследований» (АРСО-2001). — М., 2001. — с. 110–116 [Электронный ресурс] http://expert.com.ua/kniise/articles/zhel1201_2.htm

Всесторонность и полнота экспертного исследования подразумевает изучение речевого следа во всех идентификационно значимых аспектах, определяющих индивидуальность и неповторимость функционально-динамических навыков звучащей речи проверяемого лица. В этой связи для того чтобы оценить, влияние искажений на пригодность речевого сигнала для идентификации, необходимо знать: тип и марку мобильного телефона, используемого абонентами при разговорах, в какой сети разговоры происходили, какой кодек и с какой скоростью передачи речевого сигнала использовался, какое устройство применялось для регистрации сигнала на материальном носителе. Без такой информации эксперт не сможет не только провести всестороннее и полное исследование речи, но даже достоверно оценить пригодность фонограмм для идентификационного исследования.

Придание большей значимости результатам акустического анализа может привести к экспертной ошибке, поскольку речь двух заведомо разных лиц, передаваемая по каналам сотовой связи, подвергшаяся разным алгоритмам цифровой обработки, по результатам акустического сравнения может случайным образом совпасть.

При отсутствии у экспертов достоверной информации о технологической цепочке записи представленных фонограмм возрастает роль собственно лингвистических признаков речевой индивидуализации.

Лингвистические идентификационные признаки, которые отражают индивидуальные речевые навыки говорящего, определяются индивидуально-стилистической манерой речи, навыками владения языка, на котором осуществляется речевое общение, а также интеллектуальными способностями и особенностями изложения мыслей. Система и классификация лингвистических признаков речевой индивидуализации для русскоязычных дикторов достаточно подробно изучена, описана, внедрена в экспертную практику[12].

Устная речевая деятельность — продукт реализации двух видов навыков, определяющих физическую и собственно языковую сторону речевого сигнала в коммуникации. Поскольку наибольшему воздействию подвержена физическая сторона сигнала, то возрастает роль устно-речевых навыков, в основе которых находятся лингво-психологические автоматизмы речевой деятельности, а также рече-двигательных навыков, базирующихся на механизме функционально-динамического стереотипа. Однако такой подход требует применения процедуры отождествления диктора, базирующейся на теоретических основаниях, отличающихся от принятых в естественнонаучных приложениях, где идентификация сводится к сравнению нескольких автоматически или полуавтоматически измеряемых параметров речевого сигнала и подразумевает вероятностное (в той или иной мере достоверное» «попадание в цель» или «пропуск цели» с учетом допускаемых пределов (ошибок первого, второго рода). Задача «верификации» (подтверждения тождества путем попарного сравнения диктора и образца его голоса) проще задачи «идентификации» (установления

---

[12] См., подробнее Галяшина Е. И. Судебная фоноскопическая экспертиза. — М.: Триада, 2001. с. 200–208.v

тождества многомерным сравнением неизвестного диктора с образцами речи нескольких подозреваемых дикторов»), т. к. при верификации принимается альтернативное решение.

Экспертная идентификация диктора по фонограмме речи скорее соотносится с задачей идентификации на открытом множестве дикторов, совмещая в себе наиболее сложные случаи верификации, требующей увеличения количества образцов дикторов в эталонном множестве, объем которого соизмерим с числом населения земного шара. При том, что при бесконечно большом числе дикторов вероятность ошибки идентификации стремится к единице.

Понятие лингвистического идентификационного признака определяется как выражение индивидуализирующего свойства системы речемыслительных навыков человека. Для того чтобы лингвистический признак мог быть использован в качестве идентификационного, он должен отвечать нескольким условиям. Главным является специфичность, оригинальность признака, который наиболее точно и полно отражает свойство объекта. Под оригинальностью понимают нетипичность, отклонение признака от средних величин и норм. Такой признак имеет тем большее значение, чем реже он встречается у однородных объектов одной группы. Под специфичностью признака понимают его способность выделять объект, отграничивать его от группы схожих, похожих объектов. Второй особенностью идентификационного признака является его выраженность, способность к систематическому адекватному отображению в речевом следе, моделе (когда приходится иметь дело не с самими признакам и, а с их отображениями) при передаче по каналам цифровой связи. Признак должен быть воспроизводим в каждом случае образования речевого следа, и его отображение должно однозначно передавать информацию о свойствах речемыслительных навыков. Третье — признак должен быть относительно устойчивым при модификации речевого сигнала. Лингвистические признаки могут быть групповыми или частными. Групповые признаки — присущи однородным объектам, они позволяют выделить данную группу из других подобных групп, а также отнести объект по отображению групповых признаков к данной группе (например, признаки говора, диалекта, социолекта, гендера и т. п.). Частный признак позволяет выделить конкретный объект из группы однородных ему.

Индивидуализировать объект способна только индивидуальная совокупность частных лингвистических признаков, неповторимая в своей комбинаторике и локализации.

Экспертная практика показывает, что семантическая структура устного текста как отражение речемыслительной деятельности человека уникальна и неповторима. Выбор языковых средств и их упорядоченность в тексте идет от субъективного личностного смысла, отношения говорящего к действительности и его коммуникативной цели.

По мере формирования механизмов речи у человека образуется функционально-динамический стереотип устной речи, основанный на множественной системе языковых и фонационных навыков. Система языковых навыков, входящих в структуру функционально-динамического комплекса имеет сложную

структуру, но в ней можно выделить совокупность речевых средств, включающих в себя как вариации трехуровневой модели языковой личности[13], так и ее коммуникативную способность, т. е. не только языковые компетенции, но и речевые умения.

На каждом уровне организации языковой личности (вербально-семантическом, лингвокогнитивном и мотивационно-прагматическом) можно выделить соответствующую подструктуру лингвистических признаков, характеризующих проявление языковой личности в конкретной ситуации речевого акта: индивидуальный лексикон, тезаурус концептов, признаки дискурсивного мышления и интеллектуальных навыков.

Лексикон личности понимается как словарный запас и динамическое устройство с подвижными связями и отношениями, которые обеспечивают оперирование языковыми единицами в процессе построения и понимания текста. Экспертная практика показывает, что в индивидуально-личностный лексикон входят помимо словаря лексических единиц и закрепленные в речевом опыте грамматические значения и навыки, определенные синтаксические шаблоны фраз с застывшей в них грамматикой[14]. Это — устойчивое употребление стереотипного набора словосочетаний и синтаксических конструкций при описании одних и тех же объектов, ситуаций, обстоятельств. Наличие речевых стереотипов указывает на ведущую роль лексико-семантической организации текста, являющейся вербализацией языковой картины мира говорящего. Особенности этой организации, выражающей вектор семантического наполнения высказывания посредством предпочитаемых личностью языковых форм и речевых оборотов, составляет тезаурус, отражающий иерархически организованные концепты. Понятийно-смысловая организация текста позволяет выявить признаки интеллектуальных и познавательных навыков человека, которые невозможно выявить только путем анализа лексико-грамматической, стилистической организации текста. В качестве интеллектуальных навыков идентификационно значимыми являются те особенности мышления личности, которые соотносимы со способом представления действительности в наглядных образах или абстрактных понятиях, т. к. по ним можно судить по соотношению конкретной и абстрактной лексики.

В качестве идентификационных признаков выделяются индивидуальные предпочтения по выбору стереотипных языковых средств (фразеологизированных и лексически ограниченных речевых оборотов), используемых для выражения эмоций, чувств, состояний человека, его отношения к действительности, поступках или поведению иных лиц, ситуаций и т. п.

---

[13] Караулов Ю. Н. Русский язык и языковая личность. — М.: УРСС, 2002, с. 60–61.

[14] По Н. Ю. Шведовой — «шаблонные фразы», не требующие «комбинирования» и не поддающиеся отчетливому членению — «построения в субстантивном значении этого слова». См. Н. Ю. Шведова Очерки по синтаксису русской разговорной речи. //Ран Институт русского языка. Отв. Редактор В. В. Виноградов. — М.: Азбуковник, 2003, с. 7.

Существуют два метода определения идентификационных лингвистических характеристик: 1) исследовательский анализ речевого потока по определенному набору признаков (например, по базе автоматизированной системы «Диалект» и ее производных) и принятие решения по результатам сравнения значений признаков по лингвистическим протоколам, 2) прямая аудиторская оценка экспертом текста с последующим соотнесением полученного лингвистического описания с набором лингвистических признаков, зафиксированных в базе автоматизированной системы, и оценкой полученных совпадений и различий.

Нами была поставлена задача выявить среди названных признаков наиболее специфичные, отвечающие критериям надежности, воспроизводимости и устойчивости. С этой целью были проанализированы признаки как плана выражения (сегментного и супрасегментного уровня), так и плана содержания (лексико-семантического и семантико-синтаксического уровня). Проведенный анализ позволил сделать следующие выводы.

Особенность языковой организации речевой реализации приобретает идентификационную значимость только в том случае, если, во-первых, данная реализация является нарушением языковой системы, во-вторых, отклонением от принятого в данной социальной общности употребления языковых средств в устной речи; в-третьих, нарушением узуса — несоответствием используемых языковых средств целям и условиям данной ситуации коммуникации, связанное с неумением выбрать те из них, которые наиболее адекватно отражают социальные нормы поведения в данном дискурсе, в-четвертых, речевая реализация является индивидуально предпочитаемым вариантом из ряда вариаций, допускаемых языковой системой и (или) узусом.

Значениями признака будут его конкретное проявление и частота встречаемости в речи данного диктора по сравнению с среднестатистической частотой встречаемости в данной языковой общности.

Наибольший идентификационный вес имеют те признаки, которые относятся к развернутым семантическим комплексам с присущими им фонетической организацией, синтаксическими и синтагматическими связями. К ним относятся особенности структурной организации понятийно-смысловых элементов устного текста и языковые построения, отображающие навыки дискурсивного мышления и интеллектуальных навыков.

К дискурсивным навыкам мышления традиционно относят: использование того или иного вида определения, деления понятия; того или иного вида силлогизма (категорического, условно-категорического, условного и т. д.), энтитемы (сокращенного силлогизма), например предпочтительное использование энтитем с пропущенной большей или меньшей посылкой или заключением; того или иного вида доказательства и т. д.; особенности общей логической структуры (формы аргументации); наличие акцентуации — той или иной формы выделения важных для говорящего понятий, суждений; характер исправлений смыслового характера. К интеллектуальным навыкам относят: признаки интеллектуальных навыков восприятия действительности (характер восприятия), характеризующий аналитический, синтетический или смешанный

тип восприятия; признаки навыков переработки информации, характеризующие описывающий, объясняющий или смешанный тип, признаки интеллектуальных навыков акцентуации и аргументации, на какие аспекты (авторитеты, факты, обстоятельства) говорящий акцентирует внимание для обоснования своего мнения, признаки навыков оценки описываемых предметов, ситуаций.[15]

К этим признакам относятся: характер выражения субъективной модальности (вводные слова и конструкции, междометия, частицы, используемые в значении акцентирования или, оценки, отражающие чувства и переживания, отношение говорящего к предмету речи или действительности); характер и способ использования когнитивных элементов, структур и форм связи между ними (характер и способ аргументации, наличие или отсутствие логики изложения, характер акцентуации), а также способ отображения характера оценки (этической, эстетической, прагматической, эмоционально-экспрессивной).

Проведенное исследование позволило сделать следующие выводы. Индивидуальные особенности лексико-семантического и семантико-синтаксического уровней речи дают возможности судить об индивидуальных навыках речемыслительной деятельности, стилистике речи, языковой и коммуникативной компетенции говорящего. Устный текст как системное целое с иерархической организацией составляющих его структур (языковых, логических) может рассматриваться как проекция языковой личности, которая включает наряду с лексиконом, тезаурус личности, где отображен образ мира с прагматикой, системой целей и мотивов, установок и ценностных ориентаций.

Языковой компонент речевого сигнала, интерпретируемый как устный текст, а, следовательно, и содержащаяся в нем информация о дикторской индивидуальности независимы от акустической составляющей. Лингвистическая информация о языковой личности диктора, заключенная в тексте, является чисто аддитивной величиной относительно соответствующей голосовой информации. Лингвистический анализ при идентификации говорящего по фонограммам речи имеет свой объект и при исследовании этого объекта может опираться на комплекс вербально-семантических, лингво-когнитивных и мотивационно-прагматических признаков, независимых от параметрических характеристик акустического сигнала. Иными словами, информация о дикторе, содержащаяся в лингвистическом компоненте речевого сигнала, дополняет информацию, содержащуюся в параметрическом представлении сигнала, независима от нее, может использоваться самостоятельно. Для установления тождества языковой личности диктора необходимо выявление индивидуального комплекса индивидуализирующих лингвистических признаков, определенных на всех элементах, составляющих структуру речи на трех уровнях гносеологической модели языковой личности.

---

[15] См. Роль человеческого фактора в языке. Язык и картина мира. — М., Наука, 1988; Леонтьев А. А. Лингвистическое моделирование речевой деятельности // Основы теории речевой деятельности. — М.: Наука, 1974; Вул С. И. Судебно-авторoведческая идентификационная экспертиза. Методические основы. — Х.: ХНИИСЭ, 2007, с. 14–16.

Литература

1. *Булгакова Е. В., Краснова Е. В.* (2014), Экспертные системы и методы идентификации диктора // Известия вузов. Приборостроение. Т. 57, № 2, с. 58–63.
2. *Галяшина Е. И.* (2001), Судебная фоноскопическая экспертиза, Триада, Москва.
3. *Галяшина Е. И.* (2003), Основы судебного речеведения, Стэнси, Москва.
4. *Караулов Ю. Н.* Русский язык и языковая личность. — М.:УРСС, 2002.
5. *Кураченкова Н. Б., Байчаров Н. В., Ермакова М. А.* (2007), Идентификация лиц по устной речи на русском языке. Методика «Диалект», Институт Криминалистики ЦСТ, Москва.
6. *Россинская Е. Р., Галяшина Е. И., Зинин А. М.* (2013), Теория судебной экспертизы, Норма, Москва.
7. *Россинская Е. Р., Галяшина Е. И.* (2014), Настольная книга судьи: судебная экспертиза, Проспект, Москва.

References

1. *Bulgakova E. V., Krasnova E. V.* (2014) Expert Systems and Methods of Speaker Identification [èxpertnyie sistemy i metody identifikatsii diktora].// Izvistia vuzov. Priborostroenie. Vol. 57, № 2, c. 58–63.
2. *Galyashina E. I.* (2001), Forensic Phonogram Investigation. [Sudebnaja Ffonoskopicheskaja èkspertiza], Triada, Moscow.
3. *Galyashina E. I.* (2003), The Fundamentals of Forensic Speech Science [Osnovy sudebnogo rechevedenija], Stènsi, Moscow.
4. *Karaulov Ju. N.* Russkij jazik I jazykovaja lichnost. URSS, Moscow,2002.
5. *Kurachenkova N. B., Bajcharov N. V., Jermakova M. A.* (2007), Oral Russian Language Speaker Identification. Method “Dialect” [Identifikatsija lits po ustnoj rechi na russkom jazyke. Metodika “Dialekt”, Institut Kriminalistiki TsST], Moscow.
6. *Rossinskaja E. R., Galyashina E. I., Zinin A. M.* (2013), The theory of forensic science. [Teorija sudebnoj èkspertizy, Norma, Moscow.
7. *Rossinskaja E. R., Galyashina E. I., Zinin A. M.* (2014), Forensic expertise: a judge concise book [Nastoljnaja kniga sudji: sudebnaja èkspertiza, Prospekt, Moscow.

СМЫСЛОВОЕ ВЫРАВНИВАНИЕ, ОСНОВАННОЕ НА ЛИНГВИСТИЧЕСКОЙ МОДЕЛИ, КАК СРЕДСТВО ИНТЕГРАЦИИ НОВОГО ЯЗЫКА В МНОГОЯЗЫЧНУЮ ЛЕКСИКО-СЕМАНТИЧЕСКУЮ БАЗУ ДАННЫХ С ИНТЕРЛИНГВОЙ

Гончарова М. Б. (maria_go@abbyy.com),
Козлова Е. А. (Helen_Koz@abbyy.com),
Пасюков А. В. (Artem_P@abbyy.com),
Гаращук Р. В. (Ruslan_G@abbyy.com),
Селегей В. П. (Vladimir_S@abbyy.com)

ABBYY, Москва, Россия

Ключевые слова: смысловое выравнивание, многоязычные лексико-семантические ресурсы, интеграция новых языков

MODEL-BASED WSA AS MEANS OF NEW LANGUAGE INTEGRATION INTO A MULTILINGUAL LEXICAL-SEMANTIC DATABASE WITH INTERLINGUA

Goncharova M. B. (maria_go@abbyy.com),
Kozlova E. A. (Helen_Koz@abbyy.com),
Pasyukov A. V. (Artem_P@abbyy.com),
Garashchuk R. V. (Ruslan_G@abbyy.com),
Selegey V. P. (Vladimir_S@abbyy.com)

ABBYY, Moscow, Russia

This paper presents a model-based approach to Word Sense Alignment (WSA) applied for new language integration within ABBYY Compreno lexical-semantic database with interlingua. Using the model, i.e. semantic and syntactic compatibility, we perform semantic-syntactic analysis with language-independent structure as a result. With the comprehensive description of core languages at our disposal, we analyze parallel resources, namely, the part of a bilingual dictionary and of a parallel corpus in a source language, and obtain a set of candidate concepts for meanings of a target

language. In this way, we accomplish WSA between the dictionary meanings and the concepts of interlingua. Once the correspondences between the meaning and the concepts of the hierarchy are established, these new meanings can be incorporated into the lexical-semantic database. The integration is fulfilled semi-automatically, i.e. at the final stage the correspondences are to be approved by a linguist; however, the amount of manual work is reduced to minimum.

Key words: word sense alignment, multilingual lexical-semantic resources, new language integration

1. Introduction

In recent years, quick integration of new languages into multilingual lexical-semantic resources (LSR) has been one of the key challenges facing the NLP-community. Despite being time and money consuming venture, the task is nevertheless indispensable for all cross-lingual NLP applications based on semantics. Initially, LSR were mainly expert-built, which required years of manual work. The most well-known and inventory-rich expert-built lexical-semantic database is Princeton WordNet (PWN) and multilingual resources centered around it.

ABBYY Compreno Technology was also created on the basis of a multilingual LSR developed by linguists. The system is centered around interlingua, a hierarchy of language-independent concepts serving as a link between languages and resources, and is based on the model. The term 'model' stands for a full description of semantic and syntactic compatibility of a given meaning [Manicheva et al., 2012]. Therefore, the description is voluminous and requires much effort in terms of manpower and duration. However, the already existing comprehensive description allows to speed up new language integration considerably.

Within the present article, we report on the approach to new language integration hinging on model-based WSA. One of the key implementations of WSA [Matuschek, 2014] is to bring together heterogeneous pieces of information pertaining to a given meaning presented in different LSRs. However, thanks to interlingua and language-independent output of semantic-syntactic analysis, WSA can also be employed for new language integration within interlingua-based systems such as Compreno.

As stated above, we define our approach as model-based. Using the model as a reference point, we perform a semantic-syntactic analysis of a part of the available bilingual resources (bilingual dictionaries and parallel corpora) in a source language that has already been described (in our case, English). Due to universal structure of Compreno LSR, the semantic-syntactic analysis provides a set of candidate language-independent concepts for the meanings of the target language (in our case, German). Once the correspondences between the meaning and the concepts of the hierarchy are established, these new meanings can be incorporated into Compreno lexical-semantic database. The integration is fulfilled semi-automatically, i.e. at the final stage the correspondences are to be approved by a linguist.

The paper is organized as follows. Section 2 presents the existing approaches to WSA and new language integration to LSRs. In Section 3, we concentrate on our background, briefly describing the Compreno language model and how it is used for semantic-syntactic analysis. Section 4 is devoted to the methodology of the present approach. Section 5 introduces the evaluation results. Finally, Section 6 contains our conclusions and illustrates possible further development.

2. Overview

2.1. Approaches to new language integration to LSRs

For the systems based on one core-language, we can distinguish two approaches to integration of new languages [Vossen, 1998]:

- **the merge model** presupposes creating a new hierarchy for the target language with subsequent linking of its nodes with those of the source LSR. This model was mostly used at the early stages of multilingual LSR development [Azarova, 2008; Tufis et al., 2004].
- **the expand model**: The expand model exploits the structure of PWN filling it with the meanings of new languages[Pianta et al., 2002, Robkop et al., 2010; Wang and Bond, 2013]. Being mostly translation-driven, this model relies on various bilingual [Oliver and Climent, 2014; Pradet et al., 2014; Fisher and Sagot, 2008] as well as collaboratively-built (Wikipedia, OmegaWiki, Wiktionary) resources [Pilehvar and Navigli, 2014].

The expand model approach to integration of a new language on the basis of parallel bilingual corpora originates from the presumption that the translations of words in real texts shed light on their semantics [Resnik and Yarowsky, 1997; Mikolov et al., 2013]. There are two strategies for automatic construction of such corpora:

- **by machine translation of sense-tagged corpora** [Oliver and Climent, 2012]
- **by automatic sense tagging of bilingual corpora [**Oliver and Climent, 2014].

The same methods can be applied to interlingua-based LSRs, as our current work demonstrates. As a matter of fact, in our experiments we are using a set of methods associated with the expand model because we process various bilingual resources.

2.2. Approaches to WSA

The primary goal of WSA is to unify the information associated with a given meaning through linking pairs of senses (or, more generally, concepts) from two LSRs, where the members of each pair represent an equivalent meaning [Matuschek, 2014]. There are several approaches commonly used for this task: approaches based on the similarity of textual descriptions of word senses, approaches based on structural properties of LSRs, and a combination of both.

In the framework of **similarity-based approaches,** the meanings are aligned according the similarity glosses, i.e. textual descriptions of word senses. Using this method, Niemann and Gurevych [2011] aligned WordNet to Wikipedia, while Meyer and Gurevych [2011] aligned WordNet to Wiktionary, calculating cosine or personalized page rank (PPR) similarity [Agirre and Soroa, 2009] and using simple machine learning techniques for sense classification. Later on, the same approach was chosen for cross-lingual alignment between WordNet and the German part of OmegaWiki [Gurevych et al., 2012], with machine translation as an intermediate component.

Within **graph-based approaches,** structural properties of LSRs are the main criteria for linking senses. Thus, Ponzetto and Navigli [2009] built subgraphs of WordNet for each Wikipedia category to align WordNet synsets and Wikipedia categories. Alternatively, Matuschek and Gurevych [2013] apply a kind of graph-based approach, Dijkstra-WSA, to align different resources (WordNet-OmegaWiki, WordNet-Wiktionary, GermaNet-Wiktionary and WordNet-Wikipedia) using the shortest path lengths.

Currently, a **hybrid approach** is also in use, where distances between senses in the graph representations of LSRs are taken into account along with gloss similarities [Matuschek and Gurevych, 2014].

As we have already pointed out, in this paper we present a model-based approach to WSA. We align the German meanings in a bilingual dictionary with the concepts of the SH through parsing of a bilingual German-English dictionary and a parallel German-English corpus. A more detailed description of Compreno semantic model and the process of semantic-syntactic analysis will help to understand how this approach was developed.

3. Background

3.1. Compreno Description

The core of Compreno linguistic model is a universal **Semantic Hierarchy** (SH) based on interlingua. **Interlingua**, a language-independent level of concepts, serves as a link between different languages (Fig.1). At present, the SH contains 141,342 concepts. The description of Russian and English are almost complete; the integration of German is well underway; French, Chinese and Spanish are at the initial stage of description.

The SH is a hierarchical tree organized according to hyper-hyponymy relations. For each node only one direct ascendant is possible. The nodes of this tree are called **Semantic Classes** (SC) and represent language-independent "meanings". An SC contains **Lexical Classes** (LC) that represent language-specific meanings. In their turn, the LCs contain words, i.e. language-specific lexemes. It is worth mentioning that the structure of Compreno SH is POS-independent; consequently, lexemes belonging to different parts of speech can be comprised within a single meaning, depending on the model of the branch. The meanings within SC are either synonyms, or antonyms, and differ by a set of **semantemes**, units of universal semantic information, e.g. <<PolarityPlus>>, <<PolarityMinus>>, <<Bookish>>, <<Special>>,

<<Elevated>>, etc. Semantemes also encode more specific semantic relations that are not explicitly reflected in the structure of the SH, for instance, <<Part>>, <<Whole>>, or <<SingulativePortion>>.

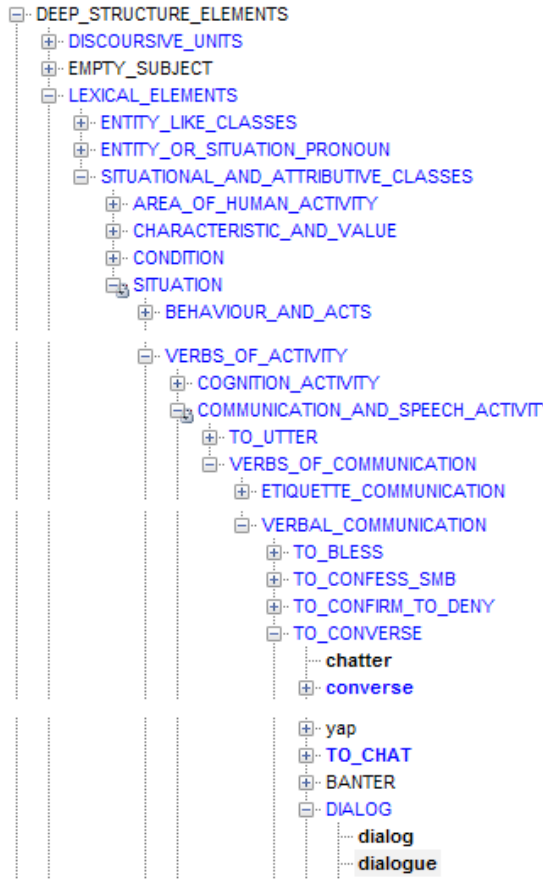


Figure 1. A Fragment of the Semantic Hierarchy

Since from the very beginning the system has been conceived as a multilingual database aimed at machine translation, each meaning is provided with a morphological, lexical semantic, and syntactic description.

The key feature of Compreno technology is that each concept and each meaning in the SH has **a semantic and syntactic model**, i.e. semantic and syntactic compatibility, which is inherited from the higher levels of the SH. Semantic compatibility is described by means of language-independent **semantic slots** (more than 300), which, to some extent, correlate with semantic valencies in L. Tesnière's dependency grammar theory [Tesnière, 1959], with deep cases in Ch. Fillmore's case grammar theory [Fillmore, 1968]. Syntactic compatibility, on the other hand, is described with the help

of **syntactic slots** that represent language-specific realizations of semantic slots. Syntactic characteristics of a meaning are unified within a **syntactic paradigm,** which includes a **universal syntactic paradigm** (syntactic characteristics of different POS) and a **lexical syntactic paradigm** (syntactic properties of a given meaning). The description also comprises non-tree syntax (regulates conjunction links, structural control, pronoun resolution, etc.), and analysis rules (preserve/extract universally-relevant bits of grammatical meaning, such as Tense and Modality of verbs, or the Number of substantives).

The set of semantic and syntactic properties, coupled with unsupervised machine learning through the use of an automatically labeled corpus, allows to deal with Word Sense Disambiguation (WSD) for a concrete language. Since Compreno has been conceived as a multilingual model, it also provides features for **treatment of cross-language asymmetry**. Cross-language hyperonym-hyponym asymmetry is neutralized by the ability to choose translation equivalents from both parent and child SCs [Manicheva, 2012]. Lexical gaps are filled with multiword expressions (terms and idioms). Within our system, **terms** are not just concepts relating to a certain domain. They are always multiword and are situated right under the SC of their root nodes, inheriting all their properties. In linguistics, **idioms** are usually presumed to be figures of speech that contradict the principle of compositionality. This principle states that the meaning of a whole should be constructed from the meanings of the parts that make up the whole). In the framework of our system, idioms are positioned according to the meaning of the whole expression.

3.2. Stages of text analysis

The distinctive feature of the approach presented in the article is full semantic-syntactic analysis of bilingual resources. We perform automatic sense-tagging of the English part of the parallel corpus by means of ABBYY Compreno parser. An important aspect of Compreno parsing technology is that syntactic and semantic disambiguation are processed in parallel from the very beginning (in contrast to the architecture more usual for the NLP systems where the semantic analysis follows the syntactic one [Anisimovich et al., 2012]).

The analysis is performed in several stages (Fig. 2). Semantic ambiguity remains unresolved as long as possible. The first stage is **lexical-morphologic**, where we consider all possible LCs for a given lexeme with all possible morphological meanings. At the stage of **syntactic analysis,** we build a syntactic graph. Initially, the edges of the graph are labeled with all possible syntactic and semantic relations, as well as grammatical properties. Gradually, incompatible meanings are eliminated. At the same time, the system checks for non-tree relations, if any. As a result of the filtration of incompatible meanings, we obtain one or several semantic-syntactic structures, each of which has the right to exist due to semantic-syntactic homonymy. **The final semantic-syntactic structure** is chosen according to statistical evaluation [Zuev, 2013]. Finally, **the universal semantic structure** with semantic relations and meanings is built through removal of all the language-specific information (surface slots, LCs and grammatical meanings).

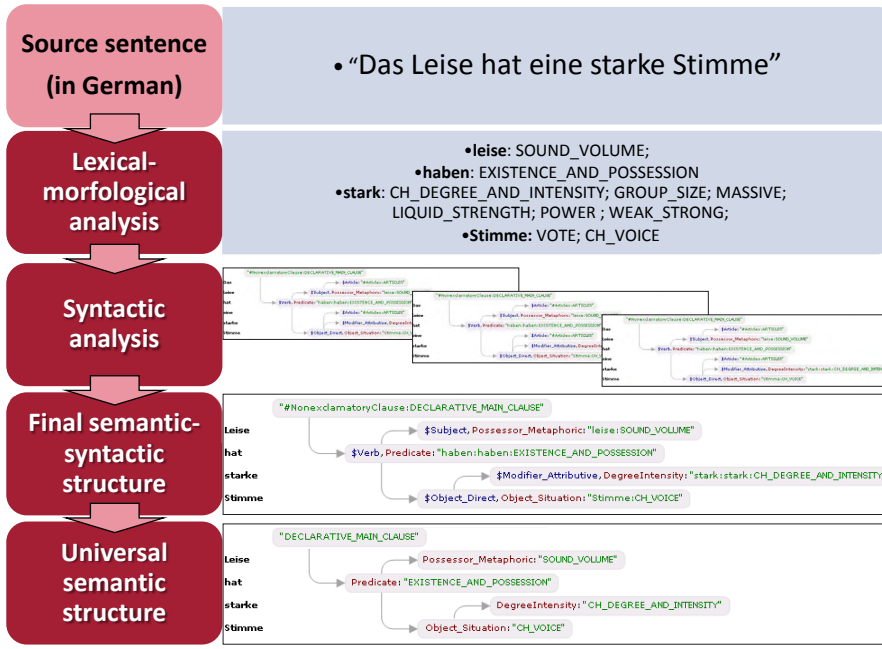


Figure 2. Stages of the Semantic-Syntactic Analysis

4. Methodology

4.1. Parallel corpus processing and statistical data retrieval

We carry out word alignment of a large parallel English-German corpus (10,250,572 sentences). Matching is accomplished using the Hungarian method for constructing a maximum weighted bipartite graph matching [Kuhn, 2010].

4.2. Statistical data filtration

As a result of word alignment, we obtain a list of pairs 'German lexeme—English translation variant'. Then, the English part of the corpus is parsed, and we obtain a list of pairs 'German lexeme—SC', with a frequency score for each correspondence. This list is called henceforth statistical data, or statistics. We filter out low frequency results (1/10000 of the maximum value) for each lexeme. At the next stage, it is important to distribute the resulting pairs across the meanings of the dictionary[1] entry.

[1] PONS Wörterbuch Englisch Premium.
Number of headwords: 98093. Number of entries: 97946. Version: 1.0 (01.11.2011)
Source: PONS Wörterbuch Englisch Premium. Based on PONS dictionary contents www.pons.de
© PONS GmbH, Stuttgart 2011

For these purposes, we perform semantic-syntactic analysis of the entry. In fact, the main principle that underlies our approach is quite simple: we obtain parallel corpus data and data from the dictionary entry and intersect the two sets (see Fig. 3).

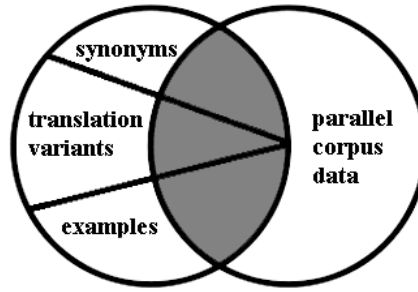


Figure 3. Venn diagram: Intersection of Data Received from the Parallel Corpus and the Dictionary

The principle described above is realized by means of a heuristically based algorithm (Fig. 4). In order to assign a given SC to the meaning of the entry, the program takes a pair 'German lexeme-SC' and decides whether it can be added through semantic analysis of the entry. We have a special environment with an integrated dictionary, where candidate SCs can be added as links to the SH. Once the link is validated by synonym, translation, or example, it is marked as a good link for the given meaning. At every stage of the analysis, the POS of the German lexeme is compared to that of the lexeme in the candidate SC. If the SC contains a lexeme belonging to the same POS, the link is marked as good. If not, the link is retained only if there are no other results.

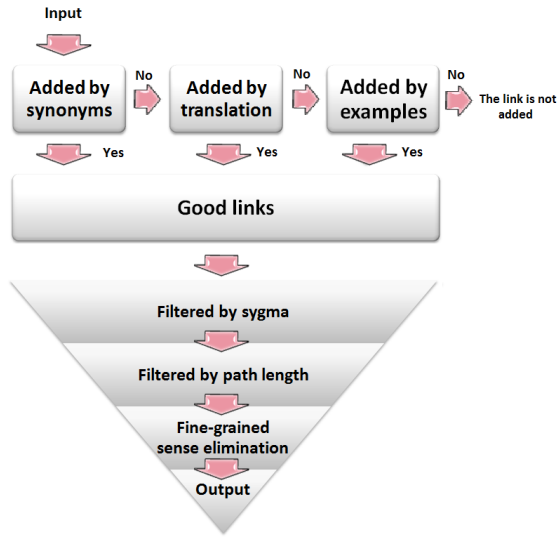


Fig. 4. The Algorithm of Adding a Candidate SC

German synonyms. As the language integration technology is semi-automatic, the German lexical semantic database is filled gradually. Consequently, we can use the analysis of those German words, which are still being added into the SH.

It often occurs that a hyperonym is indicated in brackets instead of a synonym, so we take into account parent-child relations. For example, the second meaning of *'zünden'* (Table 1) is explained through a hyperonym *'wirken'* (SC *'CH_POWER_AND_EFFECT'*). In this case, we retain SC *'TO_BLIGHT_AS_TO_AFFECT'* as a possible candidate because the class is a descendent of the SC *'CH_POWER_AND_EFFECT'*.

Table 1. Parallel Corpus, Dictionary Entry Data, and the Output for *'zünden'*

Parallel corpus data		Dictionary data	CS-candidates	Added by
{ARDENT}	114	*vt* 1) *TECH* ▪ etw zünden to fire sth *spec*	TO_BURN TO_ACTIVATE	**translation** *fire*
{DETONATION}	36			
{EMOTIONAL_STATE}	15			
{FIRE_AS_EMERGENCY}	117			
{FIRE_SHOOTING}	326			
{FIRE}	138	2) *(wirken)* to kindle enthusiasm	TO_BLIGHT_AS_TO_AFFECT	**synonym** *wirken*
{IGNITION}	21			
{INSIGHT_INTO}	30			
{TO_ACTIVATE}	225			
{TO_BLIGHT_AS_TO_AFFECT}	38	3) example: ▶ hat es bei dir endlich gezündet? — have you cottoned on? *fam*, *BRIT a.* has the penny dropped? *fam*	TO_UNDERSTAND: INSIGHT_INTO	**example** *have you cottoned on?*
{TO_BURN}	202			
{TO_COIN}	28			
{TO_CROSS_OUT}	39			
{TO_EVOKE}	21			
{TO_RETIRE}	182			
{TO_SET_THE_HOOK}	14			
{TO_TREAT_WITH_FIRE}	90			

Translation variants. When we perform semantic-syntactic analysis of the translation variants, we take into consideration all possible semantic-syntactic structures of the target language. E.g. the English translaton*'to fire smth'* for the first meaning (Table 1) gives us the SCs from statistical data*'TO_BURN'* and *'TO_ACTIVATE'* as candidate SCs.

Examples. We add the SCs from the statistics that coincide with the SCs derived from the semantic-syntactic analysis of the example, with only the best structure chosen. Thus, the SC *'INSIGHT_INTO'*, appearing in statistics, is confirmed by the example (the principle of parent-child relations is relevant here as well):

(1) #[[**have**] [**you** "#pronoun_personal:#pronoun_personal:PRONOUN_BEING"] **cottoned** "cotton_on:TO_UNDERSTAND" [**on**]]?

In this way, for each meaning we obtain a set of "good links", which undergoes a number of filtrations afterwards.

Filter 1: We calculate the standard deviation where the maximum value is used instead of the mean value (Fig. 5). Thus we determine the threshold value of frequency for every meaning. All the links that lie below the threshold are filtered out.

Filter 2. In order to reduce the number of irrelevant links, we introduce additional scores which reflect the degree of affinity with the units added by other elements of the dictionary entry for each link (Table 2). The score is calculated as a ratio between the sum of maximum coinciding path lengths and the length of a given link. To obtain the maximum coinciding path length we compare the path of a link (for instance, *'PERMISSION'*) added by one element of the entry (*'permit'*) with the links added by other elements (*'Ausweis', 'pass'*). In the example, both *'Ausweis'* and *'permit'* have the maximum coinciding path length of 6 (taken from the root of the SH). All candidate SCs with a score below 0,75 of the maximum score for a given meaning are filtered out.

Table 2. The Path Length Filter

	Ausweis	permit	pass
Legitimation *:f (geh)* **2)** *(Ausweis)* permit, pass	DEEP_STRUCTURE_ ELEMENTS : LEXICAL_ELEMENTS : ENTITY_LIKE_CLASSES : ENTITY : INFORMATION_AND_ SOCIAL_OBJECTS : CREATIVE_WORK : MATERIAL_CREATIVE_ WORK : TEXT_OBJECTS_AND_ DOCUMENTS : DOCUMENT : CERTIFICATE : Ausweis	DEEP_STRUCTURE_ELEMENTS : LEXICAL_ELEMENTS : ENTITY_LIKE_CLASSES : ENTITY : INFORMATION_AND_SOCIAL_ OBJECTS : CREATIVE_WORK : RESULTS_OF_SPEECH_MENTAL _ACTIVITY : RESULTS_OF_GIVING_INFORMA TION_AND_SPEECH_ACTIVITY : PERMISSION_PROHIBITION : PERMISSION : permit	DEEP_STRUCTURE_ELEMENTS : LEXICAL_ELEMENTS : ENTITY_LIKE_CLASSES : ENTITY : INFORMATION_AND_SOCIAL_ OBJECTS : CREATIVE_WORK : MATERIAL_CREATIVE_WORK : TEXT_OBJECTS_AND_ DOCUMENTS : DOCUMENT : WRITTEN_PERMISSION_AS_ LEGAL_DOCUMENT : PASS : pass
			DEEP_STRUCTURE_ELEMENTS : LEXICAL_ELEMENTS : SITUATIONAL_AND_ATTRIBUTIVE_ CLASSES : SITUATION : EXISTENCE_AND_POSSESSION : GIVE_GET_TAKE_AWAY : TO_GIVE : TO_GIVE_TO : pass : pass
			DEEP_STRUCTURE_ELEMENTS : LEXICAL_ELEMENTS : SITUATIONAL_AND_ATTRIBUTIVE_ CLASSES : SITUATION : POSITION_AND_MOTION : MOTION : TO_GO_AND_TRANSFER : pass : pass

Filter 3: Fine-grained sense elimination is applied when we obtain both parent and direct child SCs as candidate SCs for a given meaning. As we have already pointed out, Compreno technology has its own mechanisms for treating cross-lingual asymmetry in hyponym-hyperonym relations with certain classes marked as "transparent". Consequently, the most general concept is retained at this stage.

4.3. Additional Semantic-Syntactic Analysis of a Dictionary Entry

Additional semantic syntactic analysis of a dictionary entry is applied to multiword expressions and to dictionary meanings without candidate SCs. We extract German equivalents of English multiword expressions, irrespectively of whether their German equivalents are multiword or not (see Table 3). For definitions of terms and idioms within our system see Section 3.

Table 3. Treatment of Multiword Expressions

	English	Semantic Class	German
Terms	naval officer	NAVAL_OFFICER	Marineoffizer
	fashion designer	FASHION_DESIGNER	Modemacher
Idioms	getting rid of	GET_RID_OF	Abwicklung
	polar circle	POLAR_CIRCLE	Polarkreis

Additional semantic analysis of the entry allows us to assign links even when parallel corpus data and dictionary data do not coincide.

5. Evaluation and discussion

As a result of the semi-automatic German language integration, 121,852 meanings of 92,985 entries from PONS dictionary were assigned candidate links to SCs (Table 4).

Table 4. Number of Entries and Meanings in the Dictionary

	Meanings	Entries
Nominal	82,854	7,1808
Verbal	19,241	9,173
Adjectival	13,605	10,427
Adverbial	4,124	3,066

To evaluate the effectiveness of the method described above we have taken a random sample of 400 German lexemes from the dictionary. We established a benchmark by manually assigning correct SCs to each of these lexemes. Subsequently, we took the results of our integration method for the sample and compared the two sets. The following measures were computed:

- **precision,** that is, the percentage of relevant SCs retrieved with respect to the number of retrieved SC-candidates;
- **recall,** that is, the percentage of relevant SCs retrieved with respect to the total number of relevant manually assigned SCs.
- **F-score,** calculated according to the formula:

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

The results are presented in Table 5.

Table 5. Evaluation Results

	Overall	Monosemous words	Polysemous words
Precision	0.60	0.63	0.52
Recall	0.80	0.82	0.76
F-score	0.69	0.71	0.61

As can be seen from Table 5, we have achieved good results in terms of recall, both for polysemous and monosemous words. Since our SC-candidates are supposed to be later approved by a linguist, precision was not our primary goal. However, precision results can be still improved, which we are planning to do in the framework of our future work. In order to reduce the number of irrelevant SC-candidates, we intend to expand our use of multilingual resources.

It is common knowledge that both dictionary content and dictionary word sense distinction have a rather arbitrary and subjective nature, so it is risky to use one dictionary as a reference point; however, in our case it is checked and enriched through parallel corpus data. For sense distinction, we rely mostly on the structure of our SH, as the meanings are determined by the model and are verified by machine translation of real texts. As a result we get coarse-grained sense distinction based on empirical data.

6. Conclusion

In this article we presented a model-based approach to WSA which we use to integrate a new language (German) into Compreno lexical-semantic database with interlingua. The approach involves semantic-syntactic analysis of the English part of a parallel corpus and a bilingual dictionary. The resulting language-independent structure enables us to deal effectively with cross-language WSD and to carry out cross-language WSA of German meanings with the concepts of the hierarchy.

This approach has the following advantages. Comprehensive description of English within the system and a large-scale parallel corpus enables us to obtain a set of candidate semantic classes for practically every meaning in the German-English dictionary. There is no discrepancy between the results obtained for monosemous and for polysemous words. As the Compreno Semantic Hierarchy does not segregate words by parts of speech, we are able to process all POS in one iteration.

The interlingua level of the hierarchy can also be used for a variety of purposes, besides integration of new languages. For example, it can be applied as an intermediate component for alignment of other resources. Specifically, we are planning integration of different multilingual resources to improve our precision results. Complete replicability of the present experiment is possible within Compreno framework; replicability of the model-based approach is possible within any system with deep semantic analysis.

References

1. *Agirre E., Soroa A.* (2009), Personalizing PageRank for Word Sense Disambiguation, Proceedings of the 12th conference of the European chapter of the Association for Computational Linguistics (EACL-2009), Athens, pp. 33–41.
2. *Anisimovich K. V., Druzhkin K. Ju., Minlos F. R., Petrova M. A., Selegey V. P., Zuev K. A.* (2012), Syntactic and semantic parser based on ABBYY Compreno linguistic technologies, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog" [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog"], Bekasovo, pp. 90–103.
3. *Azarova I.* (2008), RussNet as a computer lexicon for Russian, Proceedings of the 16th International Conference Intelligent Information Systems, Zakopane, pp. 341–350.
4. *Fillmore Ch.* (1968),The case for case, in E. Bach, R. Harms (eds.), Universals in linguistic theory, New York, Holt, Rinehart and Winstonm, pp. 1–90.
5. *Gurevych I., Eckle-Kohler J., Hartmann S., Matuschek M., Meyer Ch. M., Wirth Ch.* (2012), UBY—A Large-Scale Unified Lexical-Semantic Resource Based on LMF, Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL'12), Avignon, pp. 580–590.
6. *Kuhn H. W.* (2010), The Hungarian method for the assignment problem. In 50 Years of Integer Programming 1958–2008, pp. 29–47.
7. *Manicheva E. S., Petrova M. A., Kozlova E. A., Popova T. V.* (2012), Compreno Semantic Model as Integral Framework for Multilingual Lexical Database, Proceedings of the 3rd Workshop on Cognitive Aspects of the Lexicon (CogALex-III), COLING 2012, Mumbai, 215–229.
8. *Meyer Ch. M., Gurevych I.* (2011), What Psycholinguists Know About Chemistry: Aligning Wiktionary and WordNet for Increased Domain Coverage, Proceedings of the 5th International Joint Conference on Natural Language Processing (IJCNLP), Chiang Mai, pp. 883–892.
9. *Matuschek M., Gurevych I.* (2013), Dijkstra-wsa: A graph-based approach to word sense alignment. Transactions of the Association for Computational Linguistics (TACL), pp. 151–164.
10. *Matuschek M., Gurevych I.* (2014), High Performance Word Sense Alignment by Joint Modeling of Sense Distance and Gloss Similarity. Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics, Dublin, pp. 245–256.
11. *Matuschek M.* (2014), Word Sense Alignment of Lexical Resources, Ph. D. thesis, Technische Universität Darmstadt.
12. *Mikolov T., Le Q. V., Sutskever I.* (2013), Exploiting similarities among languages for machine translation, Computation and Language Archive, abs/1309.4168.
13. *Niemann, E., Gurevych I.* (2011), The People's Web meets Linguistic Knowledge: Automatic Sense Alignment of Wikipedia and WordNet, Proceedings of the 9th International Conference on Computational Semantics, Oxford, pp. 205–214.
14. *Oliver A.* (2014), WN-Toolkit: Automatic generation of WordNets following the expand model, Proceedings of the 7th Global WordNet Conference, Tartu, pp. 7–15.

15. *Oliver A., Climent S.* (2014), Automatic creation of wordnets from parallel corpora, Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, pp. 1112–1116.
16. *Oliver A., Climent S.* (2012), Building wordnets by machine translation of sense tagged corpora, Proceedings of the Global WordNet Conference, Matsue, pp. 232–240.
17. *Pianta E., Bentivogli L., Girardi Ch.* (2002), Multiwordnet: developing an aligned multilingual database, Proceedings of the First International Conference on Global WordNet, Mysore, pp. 293–302.
18. *Pilehvar M. T., Navigli R.* (2014), A Robust Approach to Aligning Heterogeneous Lexical Resources. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014), Baltimore, pp. 468–478.
19. *Ponzetto S. P., Navigli R.* (2009), Large-Scale Taxonomy Mapping for Restructuring and Integrating Wikipedia, Proc. of the 21st International Joint Conference on Artificial Intelligence (IJCAI 2009), Pasadena, pp. 2083–2088.
20. *Pradet Q., Chalendar G. de, Desormeaux J. B.* (2014), Wonef, an improved, expanded and evaluated automatic french translation of wordnet. Proceedings of the 7th Global WordNetConference, Tartu, pp. 32–40.
21. *Resnik Ph., Yarowsky D.* (1997), A perspective on word sense disambiguation methods and their evaluation. Proceedings of the ACL-SIGLEX Workshop «Tagging Text with Lexical Semantics: Why, What, and How?», Washington, DC, pp. 79–86.
22. *Robkop K., Thoongsup S., Charoenpron Th., Sornlertlamvanich V., Isahara H.* (2010), WNMS: Connecting Distributed Wordnet in the Case of Asian WordNet, Proceedings of the 5th International Conference of the Global WordNet Association (GWC 2010), Mumbai.
23. *Sagot B., Fišer D.* (2008), Building a free French wordnet from multilingual resources, Proceedings of the Ontolex 2008, Marrakech, pp. 14–19.
24. *Tiedemann J.* (2011), Bitext Alignment, Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, United States.
25. *Tufis D., Ion R., Barbu E., Barbu V.* (2004), Cross-Lingual Validation of Multilingual Wordnets. Proceedings of the Second Global WordNet Conference, Brno, pp. 332–340.
26. *Vossen P.* (1998), EuroWordNet: a multilingual database with lexical semantic networks for EuropeanLanguages. Kluwer, Dordrecht.
27. *Wang Sh., Bond F.* (2013), Building the chinese open wordnet (cow): Starting from core synsets, Proceedings of the 11th Workshop on Asian Language Resources, a Workshop at IJCNLP-2013, Nagoya, pp. 10–18.
28. *Zuev K. A., Indenbom M. E., Judina M. V.* (2013), Statistical machine translation with linguistic language model, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog" [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog"], Bekasovo, vol. 2, pp. 164–172.

КВАНТОРНЫЕ СЛОВА, ЖЕСТИКУЛЯЦИЯ И ТОЧКА ЗРЕНИЯ[1]

Гришина Е. А. (rudi2007@yandex.ru)

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

В статье на материале Мультимедийного русского корпуса (МУРКО) анализируется жестикуляция, которая регулярно сопровождает кванторные слова со значением всеобщности *весь, все, каждый, любой*. Показано, что жестикуляция, сопровождающая кванторные слова, хорошо согласуется не столько с логическими, сколько с прагматическими особенностями кванторных слов, связанных — прежде всего — с пространственной, эвиденциальной и оперативной позицией говорящего

Ключевые слова: кванторные слова, жестикуляция, точка зрения, мультимодальность, устная коммуникация, Мультимедийный русский корпус (МУРКО)

QUANTIFIERS, GESTICULATION, AND VIEWPOINT

Grishina E. A. (rudi2007@yandex.ru)

Vinogradov Institute of Russian Language RAS, Moscow, Russia

The study analyzes gestures, which regularly accompany Russian universal quantifiers *ves'* 'the whole of', *vse* 'all', *kazhdyy* 'every', *l'uboy* 'any'. The results of the study shows that the accompanying gesticulation correlates more with the pragmatic features of the quantifiers (the spatial and evidential speaker's position and the speaker's modus operandi), than with the logical components of the quantifier's semantic structure. The Multimodal Russian corpus (MURCO) has been used as a source of the data.

Key words: quantifiers, gesture studies, viewpoint, multimodality, spoken interaction, Multimodal Russian corpus (MURCO)

---

[1] Исследование проведено при поддержке грантов РФФИ №№ 14-06-00245 и 15-06-04334, а также программы Российской академии наук «Корпусная лингвистика»

1. Введение

Данная статья посвящена обсуждению жестикуляции, которая сопровождает употребление в речи кванторных слов[2], обозначающих всеобщность. Будут рассмотрены только собственно кванторы, т. е. слова, «одна из валентностей которых соответствует некоторому объекту, а другая обозначает приписываемый ему признак и прототипически заполняется группой сказуемого» [Богуславский 2005: 144]. Таковыми словами мы считаем слова *весь* (включая *всё*), *все, каждый, любой*.

База данных, на основе которой проведено исследование, включает в себя почти 400 контекстов, отобранных из Мультимедийного русского корпуса (МУРКО): 119 контекстов с *весь*+*всё*, 90 — *все*, 84 — *каждый*, 76 — *любой*[3].

Прежде чем привести данные, следует сделать несколько **методологических замечаний.** Общеизвестно, что жестикуляционный и речевой ряды выровнены между собой, и информация в высказывании в большинстве случаев передается обоими модусами — собственно речевым и жестикуляционным. При этом, однако, существует несколько факторов, которые влияют на определенную асимметрию в расположении и кодировании информации с помощью собственно языковых и с помощью жестикуляционных средств.

1) Речевая информация упорядочена линейно и развертывается во времени последовательно, единица за единицей. Жестикуляционное кодирование устроено совершенно иначе: помимо линейной горизонтальной упорядоченности, т. е. последовательности жестикуляционных единиц во времени, жестикуляция осуществляется в трехмерном пространстве и упорядочена вертикально, т. е. **несколько** смысловых единиц могут быть переданы **одним** движением руки или головы. Таким образом, сочетание горизонтального (речевого) и вертикального (жестикуляционного) упорядочивания информации в высказывании неизбежно ведет к сдвигу речевого и жестикуляционного ряда друг относительно друга.
2) Иконический характер жестикуляции, т. е. склонность к интегральной передаче информации с помощью изобразительных жестов, в которых связь между смыслом и формой мотивирована, приводит к тому, что жест формируется несколько быстрее, чем речевое высказывание, которое характеризуется аналитическим (по сравнению с жестом) характером и конвенциональной связью между формой и смыслом. В результате жест в значительном числе случаев несколько опережает соответствующий семантический компонент в речевом потоке.

---

[2] Далее — по чисто стилистическим причинам — мы устанавливаем контекстную синонимию между словосочетанием *кванторное слово* и словом *квантор*. Поскольку в статье не используется понятие *квантор* в логико-математическом смысле, такая синонимия не приведет к путанице.

[3] Все примеры взяты из циклов телепередач «Гордон» (НТВ, 2001–2003 гг.) «Academia» (т/к «Культура), а также из видеозаписей докладов, сделанных на конференции «Диалог».

3) Точно так же, как многозначны языковые единицы, многозначными являются и жестикуляционные единицы (например, указание пальцем может обозначать определенность, отдаленность объекта указания от говорящего, императивную иллокуцию, изображение точки или траектории, эмфазу и др.). Это приводит, среди прочего, к тому, что в ряде случаев мы не можем точно сказать, какой именно семантический компонент передается данным жестом в данном контексте: например, если речевая единица характеризуется определенностью, иллокутивный акт, ее содержащий, является императивом, и сама единица подпадает под эмфазу, то трудно точно сказать, какой именно смысл передает сопровождающее указание указательным пальцем, — все значения в этом случае могут быть интегрированы в данном жесте.
4) Наконец, существует такое явление, описанное Д. Макниллом, как подхват (catchment), когда один и тот же жест осуществляется на протяжении данного высказывания, периодически прерываясь. В связи с этим бывает трудно понять — в данной конкретной точке высказывания мы имеем дело с самостоятельным жестом или с подхватом предыдущего жеста?

Как же в таком случае мы можем определить совокупность значений, которые могут быть в принципе переданы данным жестом? Единственным способом разрешить эту ситуацию является статистический анализ большого количества контекстов, которые сопровождаются данным типом жестикуляции. Принцип здесь следующий:

1) на **дескриптивном** этапе контексты упорядочиваются по речевым и жестикуляционным параметрам: контексту A приписывается ряд лингвистических характеристик (лексема, ее число, актанты и сирконстанты, тип иллокуции, время, модальность и мн. др.), а сопровождающему контекст A жесту G приписывается ряд жестикуляционных характеристик (направление движения, конфигурация руки, траектория движения, кратность и мн. др.);
2) на **дистрибутивном** этапе с помощью статистических методов (мы пользуемся методом χ-квадрат) устанавливаются статистически значимые связи между лингвистическими параметрами контекстов A и жестикуляционными параметрами жестов G по всей базе данных;
3) на **интерпретационном** этапе делается попытка объяснить те или иные полученные зависимости.

Именно благодаря такой методике мы получаем для того или иного жеста G некоторый набор его вероятных значений. Это позволяет нам при анализе каждого конкретного контекста отсекать маловероятные интерпретации сопровождающей жестикуляции и выбирать только высоковероятные. Но при этом, очевидно, все наши интерпретации конкретных контекстов будут априори иметь не абсолютный, а лишь относительный, вероятностный характер. Для того, чтобы приблизиться к абсолютному толкованию жеста, следует получить доступ не просто к **полному** набору его вероятных значений, но и к их относительному весу: например, при анализе значения того же указания указательным пальцем

в конкретном контексте, для того, чтобы точно определить его значение в этом контексте, нужно будет понять, какой фактор для данной конфигурации ладони более весом, — тип иллокуции, например, или отдаленность объекта указания. Или, быть может, эмфаза? Но для того, чтобы перейти к определению веса того или иного параметра, следует выделить если не все параметры, то хотя бы самые очевидные. Поэтому нашей задачей является не столько толкование конкретного жеста в конкретном контексте, сколько выявление основных семантических доминант для данного движения руки/головы или для данной конфигурации ладони — при полном понимании, что мы часто не можем точно сказать, какая именно из этих доминант реализована в каждой конкретной ситуации.

После этого общего методологического введения можно перейти к изложению конкретных данных по кванторным словам.

2. Основные типы жестов, сопровождающих кванторы

2.1. Квантованные жесты

Группа квантованных жестов включает в себя ручные жесты, которые тем или иным способом квантуют (делят на однородные составные части) пространство, в котором они осуществляются. Наиболее частым способом квантовать пространство является круговое движение типа ***циклоида*** (см. рис. 1), когда рука, двигаясь вдоль некоторого вектора, с помощью последовательно повторяющихся дуговых движений делит это вектор на отрезки.

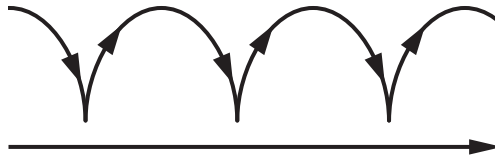


Рис. 1. Циклоида

(1) *вирусы, бактериофаги* $^{\text{циклоида}}${*вот все вот все эти вещи*} [двуручная циклоида: правая рука движется вдоль вектора вправо, левая рука — влево][4].

В укороченном варианте циклоида имеет только один цикл, тем самым оказываясь весьма близкой к движению руки по выгнутой дуге. Различие между дугой и циклоидой с одним циклом заключается в том, что при осуществлении циклоиды рука говорящего в начале и конце цикла фиксирует начальную и конечную точку с помощью движения сверху вниз.

---

[4] Далее фигурными скобками обрамляется та часть фразы, которая выровнена с ударной частью жеста и с пост-ударным удержанием, если таковое наличествует.

Квантование производится также ступенчатым перемещением открытой ладони с уровня на уровень, обычно сверху вниз — такой способ квантования используется чаще всего в контекстах, включающих компоненты ‘иерархия’, ‘уровень’, ‘ступень’, ‘этап’ и сходные:

(2) *Австралийский физиолог Бредли проследил {все стадии}* [квантованное движение открытой ладонью, ориентированной вниз, по вертикальной оси сверху вниз] *этой гибели.*

Линейное пространство квантуется также перемещением руки вдоль некоторого вектора отдельными порциями, отрезками, между которыми имеются очень короткие паузы (см. рис. 2).

Рис. 2. Отрезки

(3) *и вот {эти все аварии}* [движение обеих рук изнутри наружу сверху вниз отдельными отрезками, расположенными на двух расходящихся векторах], *которые происходят…*

Квантуется также плоскость: с помощью движения ладони, конфигурированной в *щепоть*, *кольцо* или *перо*, а также с помощью указательного пальца (см. рис. 3) в одной и той же плоскости отмечаются точки, составляющие эту плоскость.

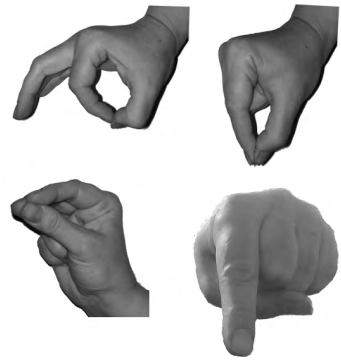


Рис. 3. Конфигурации ладони при квантовании плоскости

(4) *французской академии наук, которая в то время {тщательно следила за всеми новыми находками}* [разнонаправленное движение с фиксацией точек на плоскости, конфигурация руки *кольцо*]

Особым образом квантуется пространство, которое представляется говорящему не линейным, а объемным: в этом случае говорящий совершает разнонаправленные линейные движения двумя руками; каждое такое линейное движение символизирует собой один из радиусов данного объема и тем самым квантует объем на зоны:

(5) *взрывается, как бомба* {*во все стороны*}, *энергия* {*уходит равномерно во все стороны*} [разнонаправленные линейные движения обеими руками]

2.2. Обозначение точки

В статье [Гришина 2014b] мы писали о том, что вертикальное движение руки сверху вниз, а также вертикальный кивок могут отмечать точку на некоторой прямой. Это движение часто сопровождает конструкции с кванторными словами: рука не производит никакого движения — ни вдоль вектора, ни по окружности, а лишь движется вертикально сверху вниз, и аналогичное движение совершает голова при вертикальном кивке.

(6) *любой участок* {*любой рыбы*} [движение руки сверху вниз, конфигурация *держащая рука* (см. ниже), ориентация ладони вниз] *реагирует на свет*

(7) кивок{*каждый*} кивок{*самец*} *ранней весной, прилетая на места гнездования*

Кроме того, точка фиксируется с помощью конфигурации руки *щепоть* (см. об этом [Гришина 2014a]):

(8) *мудрость — то, что предполагает отвечать на всё, вот на* щепоть{*любой вопрос*}

(9) *древний египтянин, который умел во* щепоть{*всём*} … *видеть*

2.3. Жесты, изображающие форму объектов

Целый класс жестов, сопровождающих кванторные слова со значением всеобщности, изображает форму объектов.

2.3.1. Круговые жесты

В статье [Гришина 2015a] мы приводили корпусные данные, свидетельствующие о том, что круговые движения руками передают абстрактную идею ‘объект, имеющий форму’: поскольку окружность не имеет специально выделенных точек, таких, например, как вершины многоугольника (все точки окружности находятся на одном и том же расстоянии от ее центра), то окружность в жестикуляции передает идею формы как таковой, любой формы, т. е. передает саму идею оформленности объекта. В этом значении круговые траектории часто сопровождают кванторные слова со значением всеобщности:

(10) *и тогда симметрия нарушается* окружность{*во всем этом ареале*}

(11) окружность{*все живые*} *организмы, в отличие от неживых*

(12) окружность{*каждый*} *выглядит вороной*

(13) *если мы уберем, предположим,* окружность{*любую деталь этого куба*}.

2.3.2. Объем

С помощью одной или двух рук, ладони которых сформированы в конфигурацию *держащая рука* (см. рис. 4), говорящий может изображать объект, имеющий объем: руки/рука как бы держат в руках соответствующий объект.

Рис. 4. Держащая рука

(14) *надо тщательно анализировать* правая рука, ладонь вверх, конфигурация *держащая рука* {*каждую*} *запись*

(15) *дающее* две руки, ладони вертикально, конфигурация *держащая рука*{*всему сущему*} *существование*

Объем задается также двумя открытыми ладонями, формирующими как бы содержащий нечто контейнер (ладони образуют стенки этого контейнера, см. рис. 5).

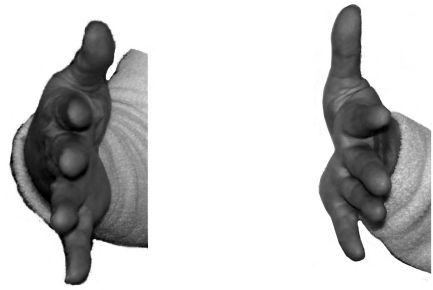


Рис. 5. Контейнер

(16) *эксперимент, в котором можно определить* две открытые ладони вертикально, движение изнутри наружу{*всю энергию*}

2.3.3. Линия

Достаточно часто форма объекта задается линией: открытая ладонь движется по модели радикальное пересечение (см. [Гришина 2013а]): поперек тела говорящего, из противоположной зоны: правая рука слева направо, левая рука справа налево (см. рис. 6).

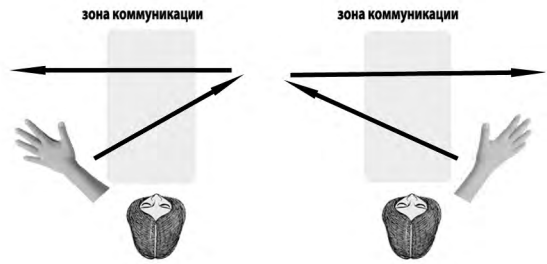


Рис. 6. Радикальное пересечение

Линия может быть также прочерчена рукой и в своей зоне (модель *комфорт*, см. рис. 7), а также может быть изображена двумя руками, движущимися изнутри наружу из единого центра, который находится в зоне коммуникации перед говорящим (данный центр может быть зафиксирован касанием ладоней в начале движения).

Рис. 7. Комфортное движение

Все эти типы траекторий относятся к множествам объектов, имеющих в качестве основного параметра формы длину — это может быть временная ось, ряд расположенных вдоль некоторой прямой объектов, объект, разворачивающийся вдоль пространственной или временной оси, и под.:

(17) *ты не можешь пойти дальше, пока ты не прошел* правая ладонь вверх, радикальное пересечение слева направо *{всего этого пути}*...

(18) *которые проявляются во всём,* правая ладонь вертикально, движение сверху вниз *{по всей} иерархической лестнице*

(19) *скорость <…> поддерживается на протяжении* правая ладонь вертикально, радикальное пересечение слева направо *{всей жизни}*

(20) *на протяжении* две вертикальные открытые ладони, движение изнутри наружу, начинается с касания в центре *{всех} этих 375 клипов*

2.3.4. Поверхность

Форма объекта может быть охарактеризована со стороны поверхности, т. е. объект, который имеет форму, в обязательном порядке имеет и некую поверхность. Идея поверхности, как было показано в работе [Гришина 2012] об ориентации ладони в указательных жестах, чаще всего передается ориентацией ладони вниз, т. е. ладонь, обращенная вниз, как бы иконически задает некую поверхность, расположенную под этой ладонью. Изредка, однако, поверхность задается и распрямленной ладонью, обращенной вверх: изображается некий большой объект, нижняя плоскость которого как бы соприкасается с обращенной вверх ладонью. В обоих случаях, и при ладони, обращенной вниз, и при ладони, обращенной вверх, идея поверхности передается дополнительно также горизонтальным движением руки. Двойная передача идеи поверхности — ориентацией ладони и горизонтальным движением руки/рук — предоставляет говорящему возможность одновременно изображать поверхность с помощью движения и передавать дополнительные компоненты смысла с помощью конфигурации руки:

(21) *это будет* конфигурация *держащая рука*, ориентация ладони вниз, рука движется слева направо, а ладонь делает многократные движения сверху вниз *{вся сплошная поверхность}, сплошь покрытая ударами*

В примере (21) движение руки передает идею поверхности, а конфигурация ладони *держащая рука*, обращенная вниз и совершающая многократные движения сверху вниз, передает идею ударов, которые астероиды наносят по этой поверхности.

2.4. Жесты отрицания

2.4.1. Качать головой

Стандартным отрицательным жестом, сопровождающим кванторы всеобщности, является жест *качать головой*:

(22) качать головой *{все} вспоминают свои студенческие годы*

(23) *меняется* качать головой *{вся картина мира}*

(24) *что отличает человека* качать головой *{от любого другого} животного*

(25) качать головой, ручной жест ‘поверхность’ *{каждый}, кому это интересно, может…*

2.4.2. Ручные отрицательные жесты

2.4.2.1. Жест типа 'выброс'

Ручные отрицательные жесты делятся на две группы. Прежде всего, это жесты, в качестве этимона которых выступает идея выбрасывания чего-либо ненужного из некоего контейнера, вместилища, расположенного непосредственно перед говорящим. Этот класс жестов включает в себя чаще всего движения открытой ладонью типа *комфорт* (см. рис. 9), т. е. левой руки налево или правой руки направо, или двумя ладонями одновременно, — и эти движения очень часто сопровождаются отбрасывающим движением кисти (К. Мюллер включает эту группу жестов, а также группу жестов 'смести с поверхности', см. следующий раздел) в Away Gesture Family (см. [Müller, Bressem 2014]), т. е. в семейство 'отбрасывающих' жестов).

(26) *если бы дали триллион, то он* 'выброс', правая рука направо{*все бы задачи*} *решил*

(27) *в отличие* 'выброс', правая рука налево{*от всех обычных*}, *даже получаемых в лаборатории*

(28) *я* 'выброс', две руки направо{*всё и отдал*} *туда*

(29) 'выброс', левая рука налево{*каждый*} *организм существует только в своей среде обитания*

(30) *должна быть сема первого лица* 'выброс', правая рука направо{*в любой форме*}

2.4.2.2. Жест типа 'смести с поверхности'

Второй группой отрицательных жестов являются жесты, которые имитируют резкое сбрасывание неких объектов с поверхности. В подавляющем большинстве случаев этот отрицательный жест осуществляется открытой ладонью, ориентированной вниз, в режиме *радикальное пересечение* (т. е. правая рука слева направо с пересечением зоны коммуникации, или левая рука справа налево с пересечением зоны коммуникации), а также двумя открытыми ладонями, ориентированными вниз, при том что каждая из рук совершает движение типа *радикальное пересечение*:

(31) *я его* 'смести с поверхности', обе ладони вниз, движение изнутри наружу{*весь разобрал*}

(32) *если <…> вырубят* 'смести с поверхности', радикальное пересечение, правая ладони вниз{*все леса*}

(33) 'смести с поверхности', радикальное пересечение, правая ладонь вниз{*в любой*} *ситуации он выживет*

3. Жесты и лексемы

Табл. 1 демонстрирует, как распределены разные группы ручных жестов между четырьмя кванторными словами со значением всеобщности (*весь, все, каждый, любой*).

Таблица 1

Группы жестов / Лексемы	Отрицательные жесты	Жесты, обозначающие форму	Фиксация точки	Квантованные жесты
весь	*4*	**72**	*1*	*0*
все	12	40	*0*	**11**
каждый	*1*	*17*	**27**	**12**
любой	**24**	*9*	3	0
$\chi^2 = 184{,}26$; $p = 6{,}52-35$, параметры связаны, распределения достоверны				

Итак, мы видим, что

- для квантора *весь* характерны жесты, обозначающие **форму**, и не характерны остальные типы жестов
- для квантора *все* характерны **квантованные** жесты и не характерна фиксация точки
- для квантора *каждый* характерны **фиксация точки** и **квантованные** жесты и не характерны отрицательные жесты и жесты, обозначающие форму
- для квантора *любой* характерны **отрицательные** жесты

В табл. 2 зафиксированы распределения жестов головы между кванторами.

Таблица 2

Движения головы / Лексемы	Качать головой	Кивок	Кивки головой вбок
весь, все	**33**	*3*	*2*
каждый	*1*	**13**	7
любой	22	*2*	**22**
$\chi^2 = 61{,}3$; $p \leq 1{,}56-12$, параметры связаны, распределения достоверны			

Мы видим, что

- для кванторов *весь, все* характерен отрицательный жест ***качать головой,*** и не характерны остальные движения головы
- для квантора *каждый* характерен вертикальный **кивок** сверху вниз и не характерен жест ***качать головой***
- для квантора *любой* характерны **боковые кивки** (собственно боковой кивок, рис. 8, и поворот головы, рис. 9), и не характерен **кивок** сверху вниз

Рис. 8. Боковой кивок

Рис. 9. Поворот головы

4. Интерпретация полученных результатов

4.1. Семантика кванторов

На первый взгляд полученные результаты выглядят достаточно случайными и слабо упорядоченными. В этом разделе мы попытаемся предложить их интерпретацию, которая, на наш взгляд, будет иметь некоторую внутреннюю логику.

Прежде всего, обратим внимание на то, что группа кванторных слов, послужившая материалом исследования, не является однородной.

Во-первых, квантор *весь* противопоставлен остальным трем кванторам, поскольку оперирует множеством как целым: кванторы *все, каждый, любой* трактуют множество как дискретное, состоящее из отдельных элементов.

Во-вторых, кванторы *весь, все* противопоставлены кванторам *каждый, любой* как **базовые** кванторы — **импликативным**. Имеется в виду следующее. В базовых кванторах идея всеобщности выражена непосредственно, без каких бы то ни было промежуточных логических ступеней. Напротив, в импликативных кванторах *каждый, любой* говорящий передает идею всеобщности через импликацию:

- ***каждый***: говорящий перебрал все элементы множества М, проверил поочередно каждый элемент множества на наличие свойства Р, установил, что каждый элемент М обладает свойством Р[5] и по результатам своей проверки **делает вывод**: все элементы М обладают свойством Р;

[5] Тем самым мы здесь не вполне согласны с утверждением Ю. И. Левина о том, что местоимения *всякий, каждый, любой* указывают на произвольный объект множества ([Левин 1973]): это верно для *всякий* и *любой*, но, с нашей точки зрения, неверно для *каждый*, который предполагает не произвольный выбор, а исчерпывающий систематический перебор.

- ***любой***: говорящий видит перед собой все элементы множества М и утверждает, что может выбрать случайным образом любой, неважно какой элемент множества М, и этот случайно выбранный элемент в обязательном порядке будет обладать свойством Р; в результате этого своего убеждения говорящий **делает вывод**: все элементы М обладают свойством Р.

В-третьих, квантор *любой* противопоставлен остальным кванторам данной группы как ирреальный[6] квантор реальным. Обратим внимание, что три квантора легко сочетаются с отрицанием (*весь мир — не весь мир, все люди — не все люди, каждый человек — не каждый человек*), в то время как *любой* испытывает существенные затруднения при сочетании с отрицательной частицей: *любой человек — *не любой человек*[7] (**не любой человек смертен*). Это обозначает, по-видимому, что квантор *любой* включает в свое толкование импликацию ‘если…, то…’: ‘если я выберу случайным образом объект из множества М, то этот объект окажется обладающим свойством Р’.

Табл. 3 подытоживает предложенные противопоставления.

Таблица 3

Свойства / Лексемы	Дискретность множества (дискретность –, целостность +)	Базовый или импликативный квантор (базовый –, импликативный +)	Реальный или ирреальный квантор (реальный –, ирреальный +)
весь	+	–	–
все	–	–	–
каждый	–	+	–
любой	–	+	+

4.2. Жестикуляция и точка зрения

Предложенные выше параметры, различающие четыре квантора всеобщности, в той или иной степени апеллируют к определенной пространственной позиции или действиям говорящего. В жестикуляционной лингвистике позиция говорящего некоторым образом неизбежно задана: поскольку жестикуляция осуществляется в пространстве вокруг говорящего, то передаваемая жестикуляцией информация так или иначе определенным образом располагается

---

[6] Можно было бы использовать термин невериди кативный квантор (ср. [Падучева 2014]) или квантор со снятой утвердительностью, но эти два обозначения показались нам слишком громоздкими.

[7] Такое сочетание возможно в поляризованных контекстах, в которых, к тому же, *любой* используется в режиме цитирования: *Или они привыкли, что за деньги любой человек на всё согласен? Ну, а теперь узнают, что не любой и не на всё!* [Алексей Слаповский. Большая Книга Перемен // «Волга», 2010]. В Национальном корпусе русского зафиксировано лишь 134 случая таких контекстов.

в этом пространстве, тем самым фиксируя как пространственное расположение говорящего, так и действия, которые говорящий метафорически совершает для передачи данного смысла. Эти общие замечания имеют непосредственное отношение к нашему материалу.

4.2.1. Квантор *весь*

Целостный, недискретный квантор всеобщности, характеризует объект как целое. С точки зрения пространственной позиции говорящего относительно объекта это означает, что говорящий находится вне данного множества, никак не может характеризовать его внутренний состав, т. е. его элементы (в связи с этим элементы данного множества могут быть как однородными — *все человечество*, — так и разнородными, — *весь дом*, — для функционирования этого квантора в речи данный параметр незначим): для внешнего наблюдателя внутреннее устройство множества недоступно, а следовательно, несущественно. Однако множество, наблюденное говорящим извне, как целое, безусловно, имеет форму, сколь угодно абстрактную, но имеет: внеположенная объекту позиция наблюдателя неизбежно порождает потребность в оформленности наблюдаемого объекта. Следовательно, вполне естественно, что при жестикуляционном обозначении множества как целого с точки зрения внешнего наблюдателя на первый план выходит единственная доступная наблюдению особенность этого множества, а именно, его форма. Это естественным образом отражено в том, что, как мы видели в табл. 1, — только квантор *весь* оценивает объект с точки зрения его формы, и именно жестикуляционная передача формы множества как целостного объекта является доминантой этого квантора. Для остальных кванторов жестикуляционное обозначение формы объекта, т. е. взгляд на множество извне, либо является незначимым параметром (квантор *все*), либо в высшей степени не характерно (кванторы *каждый* и *любой*).

4.2.2. Квантор *все*

В противоположность квантору *весь*, помещает наблюдателя внутрь множества[8]. Изнутри множества наблюдатель не может оценить форму последнего, но зато может фиксировать наличие в нем ряда отдельных элементов. Жестикуляционная передача отдельных элементов множества, его квантов, иконически естественно обозначается квантованными жестами, что и демонстрирует нам табл. 1, где квантор *все* показывает отчетливое тяготение к квантованным жестам.

Таким образом, как видим, согласно жестикуляционным данным, кванторы *весь* и *все* являются разными лексемами, с разными семантическими доминантами (целостность vs. дискретность, внешняя vs. внутренняя позиция наблюдателя относительно множества).

---

[8] Таким образом, будучи фактически тождественны с логической точки зрения («слово *весь* (*все*) указывает на то, что референт ИГ, содержащей это слово, совпадает с исходным множеством (дискретным или недискретным), т. е. включает все собственные подмножества этого множества» [Булыгина, Шмелев 1988]), *весь* и *все* различаются позицией говорящего относительно ИГ. И именно последний фактор является существенным для жестикуляционного оформления кванторных именных групп.

4.2.3. Квантор *каждый*

Квантованные жесты характерны также для квантора *каждый*, и это естественно, поскольку этот квантор предполагает последовательный и полный перебор всех элементов данного множества. Однако, в отличие от говорящего-наблюдателя в случае квантора *все*, при использовании квантора *каждый* говорящий является не только *наблюдателем* отдельных квантов данного множества, но и *оператором*, который осуществляет не просто наблюдение и констатацию наличия отдельных элементов множества, но и проверяет каждый элемент на наличие у последнего свойства Р. Таким образом, говорящий-оператор в случае квантора *каждый* в каждый отдельный момент времени имеет дело с одним-единственным элементом множества, который и подвергается операции проверки. И именно эту семантическую доминанту ‘отдельный элемент множества в каждый отдельный момент времени’ передает движение сверху вниз, фиксирующее отдельную точку на прямой или на плоскости (добавим, что именно этот семантический компонент передается также кивком сверху вниз, который в высшей степени характерен именно для квантора *каждый*, и в разной степени не характерен для остальных кванторов, см. табл. 2).

4.2.4. Квантор *любой*

Точно так же, как в кванторах *все* и *каждый*, говорящий-наблюдатель при использовании квантора *любой* наблюдает множество изнутри. Но, в отличие от квантора *все*, от говорящего в данном случае ожидается не только наблюдение, но и операция над элементом множества (случайный выбор отдельного элемента). В отличие же от квантора *каждый*, данная операция имеет не реальный, а потенциальный, виртуальный характер (‘**если бы** оператор выбрал случайно один из элементов данного множества, то **оказалось бы**, что этот элемент имеет свойство Р’). Таким образом, естественно, что для потенциального говорящего-оператора квантора *любой* ни квантование множества на отдельные элементы, характерное для *все* и *каждый*, ни реальная практическая операция над отдельным элементом множества, характерная для *каждый*, не является актуальной. Как следствие, на первый план выходит специфический компонент значения квантора *любой*, отсутствующий у всех остальных кванторов группы, а именно, ‘**не имеет значения**, какой из элементов множества М будет мной выбран, (он обязательно будет иметь свойство Р)’. Именно компонент с отрицанием ‘не имеет значения’ приводит к тому, что в контекстах с квантором *любой* доминируют отрицательные жесты типа ‘выброс’ и типа ‘смести с поверхности’.

Обратим внимание на то, что семантический компонент с отрицанием в составе квантора *любой* оформлен третьим лицом предиката (‘не имеет значения’) и фигурирует в составе ирреальной модальности (‘какой бы элемент говорящий ни выбрал’), а следовательно, отрицание здесь функционирует в ситуации максимальной коммуникативной и эпистемической дистанции между говорящим и его высказыванием. Как мы показали в [Гришина 2015b], для таких квази-объективных отрицательных конструкций характерен квази-объективный отрицательный жест рукой, а не отрицательное качание головой. И именно это мы наблюдаем

в табл. 1 в отношении квантора *любой*: его сопровождает отрицательный жест рукой, а отрицательный головной жест (см. табл. 2) для него незначим.

Напротив, для кванторов *весь, все* характерен именно отрицательный жест *качать головой*, который в положительных контекстах передает **прагматическое отрицание**, основным значением которого является высокая оценка. Такая несколько неожиданная, на первый взгляд, трансформация головного отрицательно жеста от отрицания к высокой оценке основана на том, что прагматическое отрицание, выражаемое качанием головы, содержит в себе компонент 'это удивительно, я не верю собственным глазам, ушам и под.' (см. подробнее [Гришина 2015b]). Именно компонент недоверия собственным органам чувств ведет к использованию жеста *качать головой*. Таким образом, жест *качать головой*, сопровождающий кванторы *весь* и *все*, мы рассматриваем как жест положительной оценки:

(34) качать головой{*какая неожиданная и во всех смыслах приятная*} *для меня встреча*!

Этот же жест в этом же значении сопровождает интенсификаторы высшей или просто высокой степени проявления качества — такие, как *абсолютно, совершенно, вовсе, чрезвычайно* и нек. др.:

(35) *возможность делать то, что когда-то было* качать головой{*абсолютно*} *немыслимо*

Что касается боковых кивков (см. табл. 3), то они, как мы писали в [Гришина 2013b], отсылают слушателя к невидимой зоне, находящейся за спиной говорящего, то есть выводят объект, обозначенный квантором, из зоны коммуникации в невидимую зону, расположенную за спиной говорящего. Следовательно, эти жесты головы функционируют как **головные** дублеты **ручных** отрицательных (выбрасывающих) жестов. Если это так, то это объясняет, почему данные головные жесты характерны для квантора *любой*.

5. Заключение

Кратко подытожим сказанное выше. Кванторные слова, выражающие идею всеобщности, — *весь, все, каждый, любой,* — могут быть исследованы с чисто логических и чисто лингвистических позиций, что позволяет вычленить в них общие и различающиеся семантические зоны (см. базовые работы, в которых так или иначе затронут этот вопрос: [Левин 1973], [Зализняк, Падучева 1974/1997], [Кронгауз 1984/1997], [Булыгина, Шмелев 1988], [Падучева 2003]). Выясняется, однако, что жестикуляционная лингвистика как один из активно развивающихся в последние десятилетия разделов когнитивистики может добавить к результатам логических и лингвистических исследований нечто новое. Это новое связано, прежде всего, с тем, что жестикуляция по самому своему существу не может избежать фиксации точки зрения говорящего на содержание его речи: любой жест обозначает как минимум пространственную позицию говорящего относительно той ситуации, которая отражена в высказывании.

Эта неизбежность привела к тому, что Дэвидом Макниллом в активный научный обиход было введено различение жестов наблюдателя (OVPT, observer viewpoint gestures) и жестов персонажа (CVPT, character viewpoint gestures, [McNeill 1992]). Осуществляя жест с точки зрения персонажа, говорящий как бы совмещает свое тело с телом персонажа и жестикулирует так, как будто является этим персонажем. Выбор между жестом персонажа и жестом наблюдателя диктуется, среди прочего, выбором между нарративным и диалогическим дискурсом: для нарратива пантомимные жесты персонажа характерны в большей степени, чем для диалога (они функционируют аналогично настоящему историческому времени). Жесты наблюдателя, при гораздо большей частоте использования, имеют более абстрактный характер, и для определения их значения требуются иногда обширные специальные исследования. Именно поэтому они в меньшей степени ориентированы на передачу информации слушающему и используются, скорее, как организующее средство для когнитивных процессов, происходящих в мозгу говорящего.

Анализ кванторных слов со значением всеобщности в аспекте сопровождающей жестикуляции показал, что точка зрения может оказаться **имманентным** свойством ряда лексем, не зависеть от типа дискурса и с достаточной степенью регулярности проявляться в жесте. Кроме того, помимо точки зрения, в структуру значения этих лексем может быть включен определенный modus operandi говорящего. Итоговая табл. 4 показывает, как это выглядит в отношении кванторных слов со значением всеобщности.

Таблица 4

Кванторное слово	**Пространственное расположение наблюдателя**	**Modus operandi**	**Реализация в жесте**
весь	вне множества	наблюдение за множеством как целым	ручные жесты, обозначающие форму; отрицательный жест *качать головой*
все	внутри множества	наблюдение за элементами множества	квантованные жесты; отрицательный жест *качать головой*
каждый		полный реальный перебор всех элементов множества	фиксация точки; квантованные жесты; вертикальный кивок
любой		теоретический случайный выбор одного объекта множества	ручные отрицательные жесты; боковой кивок

References

1. *Boguslavskiy I. M.* (2005), The quantifier's valency [Valentnosti kvantornykh slov], Logical analysis of language. Quantificative aspect of language [Logicheskiy analiz yazyka. Kvantifikativnyy aspect yazyka], Moscow, pp. 139–185
2. *Bulygina T. V., Shmelev A. D.* (1988), Some remarks on the words of 'some' type (Quantification in Russian) [Neskol'ko zamechaniy o slovakh tipa neskol'ko (K opisaniyu kvantifikatsii v russkom yazyke)], Language: system and functioning [Yazyk: sistema i funktsionirovanie], Moscow, pp. 44–54
3. *Grishina E. A.* (2012), Hand pointing as a system (on the data of Multimodal Russian corpus) [Ukazanie rukoy kak sistema (na materiale Mul'timediynogo russkogo korpusa), Linguistics issues [Voprosy yazykoznaniya], 3, pp. 3–50
4. *Grishina E. A.* (2013a), Gestural profiles of Russian prefixes [Zhestikulyatsionnye profili russkikh pristavok], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2013" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2013"], pp. 255–271
5. *Grishina E. A.* (2013b), Head pointing as a system [Ukazanie golovoy kak sistema], Linguistics issues [Voprosy yazykoznaniya], 3, pp. 90–130
6. *Grishina E. A.* (2014a), *Ring* and *grappolo*: fingertip connections in Russian gesticulation and their meanings [*Kol'tso* i *shchepot'*: semantika soedinennykh pal'tsev v russkoy zhestikulyatsii], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2014" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2014"], pp. 184–203
7. *Grishina E. A.* (2014b), Vertical axis in gesticulation: linguistic viewpoint [Vertikal'naya os' v zhestikulyatsii: lingvisticheskiy aspect], Russian language: scientific viewpoint [Russkiy yazyk v nauchnom osveshhenii], 1 (27), pp. 42–89
8. *Grishina E. A.* (2015a), Circles and swings: Complex trajectories and their meaning in Russian gesticulation [Krugi i kolebaniya: semantika slozhnykh traektoriy v russkoy zhestikulyatsii], Language and thought: Contemporary cognitive linguistics [Yazyk i mysl': sovremennaya kognitivnaya lingvistika], Moscow, pp. 238–286
9. *Grishina E. A.* (2015b), Russian gestures of negation [O russkom zhestikulyatsionnom otritsanii], Proceedings of Institute of Russian Language RAS [Trudy Instituta russkogo yazyka RAN] (forthcoming)
10. *Krongauz M. A.* (1984/1997), Referential type of nominal groups with the pronouns *vse, vsyakiy, kazhdyy* [Tip referentsii imennykh grupp s mestoimeniyami VSE, VSYAKIY, KAZHDYY], Semiotics and Informatics [Semiotika i informatika], 35, Moscow, pp. 227–243
11. *Levin Yu. I.* (1973). Sematics of pronouns [O semantike mestoimenij], Questions of grammatical modeling [Problemy grammaticheskogo modelirovaniya], Moscow, pp. 108–121
12. *McNeill D.* (1992), Hand and mind: What gestures reveal about thought, Univ. of Chicago Press, Chicago

13. *Müller C., Bressem J.* (2014), The family of away gestures. Embodied roots of negative assessment, refusal, and negation, Proceedins of the International conference of ISGS'2014, San-Diego, Ca, p. 21, http://isgs.ucsd.edu/files/2013/06/ISGS_Talk_Abstracts-Jul9.pdf
14. *Paducheva E. V.* (2003), Utterance, and its correlation with reality [Vyskazyvanie i ego sootnesennost' s deystvitel'nost'yu, Moscow
15. *Paducheva E. V.* (2014), Disassertion and non-veridicallity [Snyataya utverditel'nost' i neveridikativnost'], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2014" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2014"], pp. 489–507
16. *Zaliznyak A. A., Paducheva E. V.* (1974/1997), Contextual synonymy of Singular and Plural of nouns [O kontekstnoy sinonimii edinstvennogo i mnozhestvennogo chisla sushchestvitel'nykh], Semiotics and Informatics [Semiotika i informatika], 35, Moscow, pp. 7–14

ИСПОЛЬЗОВАНИЕ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ В АНАЛИЗЕ ТЕКСТОВЫХ ФОРУМОВ ДЛЯ ПОДГОТОВКИ УЧЕБНЫХ ОБЪЕКТОВ

Грозин В. А. (vlad.grozin@yandex.ru),
Добренко Н. В. (graziokisa@gmail.com),
Гусарова Н. Ф. (natfed@list.ru)

Университет ИТМО, Санкт-Петербург, Россия

Нин Тао (603096136@qq.com)

Чанчуньский научно-технологический университет,
Чанчунь, Китай

Ключевые слова: отбор признаков, глубинный анализ текстов, учебные объекты, текстовый форум, машинное обучение

THE APPLICATION OF MACHINE LEARNING METHODS FOR ANALYSIS OF TEXT FORUMS FOR CREATING LEARNING OBJECTS

Grozin V. A. (vlad.grozin@yandex.ru),
Dobrenko N. V. (graziokisa@gmail.com),
Gusarova N. F. (natfed@list.ru)

ITMO University, Saint Petersburg, Russia

Ning Tao (603096136@qq.com)

Changchun University of Science and Technology,
Changchun, China

Nowadays the concept of a learning object (LO) is widely used in preparation of educational materials. Usually, LOs are parts or fragments of previously created educational content, which is very informative and pedagogically focused. However, concerning high-dynamic branches of science and technologies LOs tend to become outdated and trivial thus losing their educative value. In this situation, specialized text forums become a valuable source of knowledge. Forums contain experience of people who actually used the technology and its features. They contain both positive and

negative experience—something that is not available from official documentation at all. However, they also contain many trivial, repeated and still irrelevant posts. Also, an expert needs to extract useful messages from text forums according to his individual learning objectives.

The paper deals with the task of automatically identifying texts potentially useful for preparation of textual educational materials within text forums. For our experiments, we have selected highly inflective languages with complex grammar and rather weak text analysis tools: French, German, Russian and Chinese (Mandarin). We have overviewed non-semantic text and social features of a text forum which indicate the suitability for creation of a textual LO. We have analyzed those features. For this purpose, we have constructed linear and non-linear models of machine learning and conducted feature selection. Even for the forums providing little information about chosen topics and forums with a lot of off-topic text in dataset, these models were better than the baseline selection methods.

Keywords: feature selection, text mining, learning object, text forum, machine learning

1. Introduction

Nowadays we are facing the rapid growth of the amount of information available online, so it becomes more difficult to organize educational process according to this growth. Besides, much more persons are being involved into educational process: they are not only students and teachers, but also instructors, self-taught learners and so on, each having his/her own educational goals. They search for educational materials that satisfy their own needs.

The concept of learning object (LO) is developed for this purpose. There are different definitions for LO, the most common are [IEEE (2002), Wiley D. A., ed. (2001)]. Namely, IEEE defines LO as 'any entity, digital or non-digital, that may be used for learning, education or training'. Wiley defines LO as 'any digital resource that can be reused to support learning'.

Usually, LOs are chunks or slices of previously created educational content. Authors [Griffiths J., Stubbs G., Watkins M. (2007)] offer to develop LOs using textual course materials as a basis or source and dividing it downwards until the smallest item of information or idea is reached. Raw assets that have no inherent pedagogical aim can also be considered as lower-level LOs [Boyle, T. (2003)]

But LOs concerning high-dynamic branches of science and technologies tend to become outdated and trivial thus loosing educativeness (defined as a property that reflects the educative value of a document) [Hassan S., Mihalcea R. (2009)]. So, when seeking educational information about new technology it is often useless to refer to existing collections of LOs. On the other hand, searching the Web using one of the current search engines frequently lead to the results which badly meet the requirements of educativeness.

In this situation specialized text forums become a valuable source of knowledge. Forums contain experience of people who actually used the technology and its

features. They contain both positive and negative experience—something that is not available from official documentation at all. But they also contain a lot of trivial, repeated and still irrelevant posts. Also, expert needs to extract useful messages from text forums according to his/her individual learning objectives.

The obvious solution is to use techniques of text mining, for example text summarization or question answering systems. But the task of preparing LOs for high-dynamic branches of science and technologies has the specifics. Typical expert's question is "Is there any new (or unknown or interesting for my audience) information in this piece of text?" It is obvious that such form of a question is difficult for question answering systems. Usually, novelty consists in emergence of concepts and relations not known before; respectively the search query has to be rather wide (coarse-grained) in order to include them. It complicates application of semantic methods in text summarization. And, last but not least, the procedure of preparing LOs has to be simple and language-independent in order to be used by ordinary teachers.

In this paper, we address the task of automatically identifying information potentially useful for preparing educational materials (learning objects) within technical text forums. Specifically, we considered the non-semantic text and graph features that indicate suitability for creating textual LO (related to the chosen topic and containing detailed argumentation). Also, we examined dependence of these features' quality from forum language.

2. Related works

Information retrieval in our situation can be considered as a variant of educational data mining. It is known as a new growing research community since 1995, and different data mining techniques can be applied here [Romero C., Ventura S. (2007)]. This section reviews the related works from different dimensions: the works aiming to handle text information in online discussion boards (or forums) as well as the approaches of question answering systems.

The task of information retrieval from text forums is usually interpreted as Web Forum Thread Summarization and typically aims to give a brief statement of each thread that involving multiple dynamic topics. Traditional summarization methods are cramped here by some challenges [Ren et al. (2011)]. The first is topic drifting: as the post conversation progresses, the semantic divergence among subtopics will be widened. Besides, most posts are composed of short and elliptical messages, their language is highly informal and noisy, and traditional text representation methods have sufficient limitations here.

According to the survey in [Ren et al. (2011)], the majority of works in the area of forum summarization use extraction-based techniques [Spärck Jones K. (2007)] and single-document approach. A lot of research on automatic dialogue summarization use corpus-based and knowledge-based methods. For example, authors [Zhou L., Hovy E. (2005)] identify clusters in internet relay chats and then employ lexical and structural features to summarize each cluster. Authors [Ren et al. (2011)] propose a forum summarization algorithm that models the reply structures in a discussion

thread. In order to represent information of online forum in a learning environment authors [Carbonaro A. (2010)] uses concept-based summarization: each word in the document is labeled as a part of speech in grammar, and to handle the word sense disambiguation problem similarity measures based on WordNet is used.

Statistical methods of dialogue summarization are also of great interest for preparing LOs. For example, in [Wang U., Cardie C. (2011)] unsupervised (TF-IDF and LDA topic modeling) and supervised clustering procedures (using SVMs and MaxEnt) are used in combination for decision summarization for spoken meetings. Authors [Biyani et al. (2014)] consider the problem of extracting relevant posts from a discussion thread as a binary classification problem.

There is a number of the works devoted to multi-lingual aspects of text summarization. For example, in order to fulfill sentiment analysis of multi-lingual Web resource authors [Hogenboom A. et al. (2014)] consider English as basic and use language-specific semantic lexicons of sentiment-carrying words. Contrary to this approach, authors [Banea C., Mihalcea R., Wiebe J. (2014)] show that the multilingual model consistently outperforms the cross-lingual one. Practical experience of developing natural language processing applications for many languages is described in [Steinberger R. (2011)]. The author considers Machine Learning methods as an extremely promising approach to develop highly multilingual systems.

A fast-growing number of studies have shown that the social factor can be useful in text forum summarization regarding to educativeness. For example, authors [Li Y., Liao T., Lai Ch. (2012)] apply similar measures as used in blogs to the forums, such as counting the number of common tags and replying or citing the same threads. Authors [Yang S. J. H., Chen I. Y. L. (2008)] explain that in an online forum context a central core (strongly connected component) contains users that frequently help each other by following questioner (requester)–answerer (expert) links.

A lot of Question Answering Systems are presented in literature (see the overview [Kolomiyets O., Moens M.-F. (2011)]). They differ by models and techniques depending on the system requirements, the type of question posed, the type of interrogated data, the type of interface and other criteria. In technical forums the conversation often inquires the solution to a specific problem faced by the user and others answer by adding their experience in that field [Almahy I., Salim N. (2013)]. If answering this question is the educational objective of the LO then the approach of Question Answering Systems can be useful.

But for complex questions a deeper semantic analysis is required (broader coverage of expected answer types, semantic role labeling and discourse analysis). For example, authors [Ferrandez et al. (2009)] solve Cross-Lingual Question Answering tasks. They make a syntactic analysis of the question using a shallow parser tool and extracting the syntactic blocks of a question. Authors [Cao et al. (2011)] process on-line forums at which questions are presented in an obvious form. They aim is extract contexts and answers for them and use the structural Support Vector Machine method. However, as the analysis of literature have shown, questions of the type, which is declared in Introduction aren't processed in known Question Answering Systems.

3. Methods

For our experiments we have selected some highly inflective languages with complex grammar and rather weak text analysis tools, in particular four languages—German, French, Russian and Chinese (Mandarin). Detailed information about the forums is presented in Table 1. From each forum we allocated threads which names contained a topic of interest used in the form of a keyword. These posts' usefulness for the creation of a textual LO with an appropriate educational value was manually marked down by experts (Table 2). We have invited experts of the relevant field who are native speakers in the languages of the forums.

Table 1

#	Forum	Language	Topic	Threads/ posts	Keywords
1	gamedev.ru	Russian	Unity	10/410	unity
2	hifi-forum.de	German	Windows vs Linux	13/173	windows, linux
3	forum.modelsworld.ru	Russian	Ship modeling	3/150	ship, model
4	5500.forumactif.org	French	Ship modeling	3/150	ship, model
5	bbs.csdn.net	Chinese	cocos2d-x	11/120	cocos
6	bbs.chinaunix.net	Chinese	Linux for beginners	11/103	linux

Table 2

Scale	Comment
0	Offtopic
1	Post is on the chosen topic, but argumentation is incomplete or absent
2	Post is on the chosen topic, and the author's point of view is well-argumented with explanations or external links

Nowadays there are a lot of works proposing different features for text forums, potentially suitable for educational value evaluation [Hassan S., Mihalcea R. (2009); Biyani et al. (2012); Smine et al. (2013); Dringus, Ellis (2005); Romero et al. (2013)]. However, not all of them are suitable for machine learning due to the specifics of our task. The list of the selected characteristics is presented in Table 3. In general, the calculation procedures were created using the sources mentioned above, but with the following specifics.

We calculated text sentiment value using sentiment keywords, specific for the forum's language. The resulting values were normalized to the range from -1 (strongly negative text) to +1 (strongly positive text).

Also, simple non-semantic text features were extracted: text length, number of links and number of keywords. Keywords were chosen strictly corresponding to the name of the forum topic. A more extensive list of keywords would mean a search for synonyms and equivalents, which requires semantic analysis.

We represented social structure in the form of a social graph, where the nodes are the users, and edges indicate a link between two users. For the creation of the social graph we have used citation analysis: if person A quotes person B by explicitly mentioning his name in text, there is a guaranteed connection between A and B. We used two methods: a non-sentiment graph (edge weight is always 1) and a sentiment graph (edge weight is related to the post's sentiment value). After the creation of the graph parallel edges' weights were summed. Then, the weights of the edges were inverted [Tore Opsahl (2014)].

Node centrality is often used to find people who are important members of society. We considered some proven [Freeman L. C. (1978); White D. R., Borgatti S. P. (1994)] metric to evaluate node centrality: Betweenness centrality—the number of shortest paths between all pairs of nodes that pass through the node; inDegree—the total weight of incoming edges; outDegree—the total weight of the outgoing edges.

Position in thread is calculated as number of post in chronological order (first post has position in thread equal to one, next one is equal to two etc.).

The features selected for machine learning are listed in Table 3.

Table 3

Type	Feature	What this feature means
Post's author graph features	Betweenness, non-sentiment graph	Author's social importance
	inDegree, non-sentiment graph	How many times author was quoted
	outDegree, non-sentiment graph	How many times author quoted someone
	Betweenness, sentiment graph	Author's social importance
	inDegree, sentiment graph	With which sentiment author was quoted
	outDegree, sentiment graph	Author's quotes sentiment
Post's author features	Number of threads author is participating in	Author activity
Thread-based post features	Position in thread	Chance of off-topic
	Times quoted	Post's impact on forum
Text feature	Length	Number of arguments and length of explanations
	Links	Number of external sources/ images
	Sentiment value (calculated using sentiment keywords)	Post's usefulness
	Number of keywords	Topic conformity

The analysis was fulfilled using machine learning methods. Model creation was made in R. We chose two models: gradient boosting model in "gbm v.2.1" package

[gbm package (2014)] (for base learners we chose trees and default model parameters were used; to determine the optimal forest size we used built-in cross-validation with three folds) and linear regression (lm). Each model was trained on training dataset (so, each forum had independent models). Each model was trained on a training dataset (so, each forum had independent models). Then, models predicted grades of each message in the test set, and N most qualitative candidates were selected. After this metric NCG metric was calculated and averaged among forums. Datasets for each forum were randomly divided into training (60%) and test (40%) sets.

To calculate selection quality we used Normalized Cumulative Gain metric [Järvelin K., Kekäläinen J. (2002)]:

$$NCG = \frac{\sum_{i=1}^{N} rel_i}{NCG_{\max}^{N}}$$

Where N is number of selected posts, rel(i) is quality of i-th selected post, and is maximum possible NCG for specified N. This metric lies between 0 and 1 (assuming rel(i) is non-negative) and indicates the quality of the selection, but this metric doesn't penalize late relevant items. It is assumed that the expert will still read all selected messages, so the order does not matter.

Following methods of post selection were chosen for baselines:

- Baseline-1: Head posts in thread have special importance and often contain more useful information (due to off-topic content in later messages) [Said D., Wanas N. (2011)]. So, one method is based on selecting the first messages of each thread.
- Baseline-2: Other method is using semantic keywords list [Steinberger R. (2011)]. A broad list of topic-related keywords and synonyms in English and in the forum language were made by experts. This method selects posts with the highest number of these words. Stemming and lemmatization (package "tm" in R) were used where possible.

4. Results and discussion

Fig. 1 shows the dependence of selection method (NCG metric) on the model and N.

As one can see, linear model (lm) was the best and both models were better than both baselines. So, because our smallest forum contains 100 messages (and test set has 40), we evaluated metrics for N varying from 1 to 30.

Since our both models were better than baselines, there are good features indicating suitability for making textual LO. Our ultimate goal is to investigate which ones. For this purpose we used feature selection methods.

For analyzing linear dependencies we used significance of features. The significance is probability of observing the data assuming linearly independence of regressor and explainable variable [Gelman A., Hill J. (2006)]. If this probability is less than a certain threshold (in our case—0.05), we can reject that hypothesis and say that there is dependence between variables.

Gradient boosting model was used to analyze non-linear dependencies. After training the model (gbm is a set of trees) we can see how many times trees were divided by each variables, and estimate split efficiency. This way we will get relative information influence metric, which can be interpreted as the importance of features [Gradient (2014).].

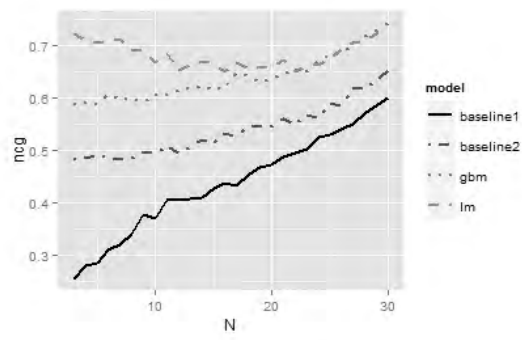


Fig. 1

Table 4 contains feature list along with their significance (S) and information importance metric (IMM) for each forum. Significance with less than 0.05 or non-zero IIM are highlighted with yellow background. Selected features (which had low significance or non-zero IIM at least once in every language) are also marked yellow. Of course, we need to consider statistical characteristics of the collected samples while selecting the features. For example: graph characteristics for forums 3 and 4 are invalid because of the small sample size and due to forum engine specifics (author names are not mentioned while quoting). Features which could not be calculated were marked as "N/A".

Table 4

	#1		#2		#3		#4		#5		#6	
Feature	**S**	**IIM**	**S**	**IIM**	**S**	**IIM**	**S**	**IIM**	**S**	**IIM**	**S**	**IIM**
Betweenness, non-sentiment graph	**0.04**	**1**	0.80	**1**	0.55	0	N/A	0	N/A	0	N/A	0
inDegree, non-sentiment graph	0.67	0	0.51	**13**	0.88	**0.7**	N/A	0	N/A	0	N/A	0
outDegree, non-sentiment graph	0.24	0	0.19	**1**	0.9	0	N/A	0	N/A	0	N/A	0
Betweenness, sentiment graph	0.17	0	0.80	0	0.57	0	N/A	0	N/A	0	N/A	0
inDegree, sentiment graph	0.67	0	0.51	**4**	0.1	**0.7**	N/A	0	N/A	0	N/A	0
outDegree, sentiment graph	0.26	0	**0.04**	**5**	0.91	0	N/A	0	N/A	0	N/A	0

Feature	#1		#2		#3		#4		#5		#6	
	S	**IIM**	**S**	**IIM**	**S**	**IIM**	**S**	**IIM**	**S**	**IIM**	**S**	**IIM**
Number of threads author is participating in	0.87	0	0.52	**4**	0.55	13.6	0.59	**17**	N/A	0	N/A	0
Position in thread	**0.02**	**4**	**0.04**	**12**	0.08	**54**	**0.04**	54	0.25	0	0.12	**14**
Times quoted	0.97	0	0.64	**6**	0.10	**1.6**	N/A	0	N/A	0	N/A	0
Length	**9e-8**	**80**	**0.03**	**42**	0.26	**15**	0.21	**49**	**0.001**	**44**	0.9	**24**
Links	0.53	0	0.97	0	0.47	**3**	0.71	0	N/A	0	0.49	0
Sentiment value	0.23	**2**	0.606	**5**	**0.001**	**22**	0.59	**39**	**1e-6**	**55**	**e-7**	**61**
Number of keywords	0.02	11	0.82	8	0.9	0	0.9	0	0.01	0	0.73	0

Posts with high educational value tend to have a positive sentiment. However, on forums ##3, 4 it has much higher significance than on forums ##1, 2, 5, 6. To explain this effect we have plotted the distribution of the number of messages from the Utility and Sentiment rounded to nearest ½. Fig 2 shows this distribution for forums ## 1, 2, 5, 6 (A) and forums ##3, 4 (B).

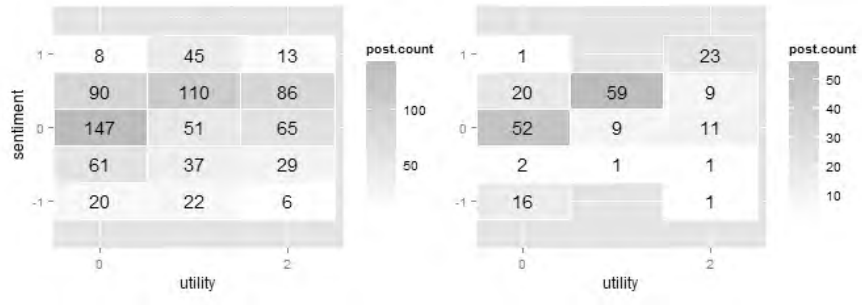


Fig. 2 (A, B)

As one can see, the distribution in the first picture (A) is nearly symmetrical with respect to sentiment=0, while the distribution in the second picture (B) is much more "skewed". For example, in fig. 2(A) value of post.count for sentiment=0.5 almost doesn't depend on value of utility. In other hand, in fig. 2(B) this dependence is obviously expressed, so a post from form B with sentiment equal to 0.5 is quite likely to have utility equal to 1. So, significance of sentiment is much higher on these forums. This "skewness" is associated with the strict moderation of forums (negative and offtopic posts are getting deleted, so users tend to leave useful and friendly texts). This is confirmed by the additional semantic analysis of the forum content, conducted by an expert.

Thus, we have selected features that indicate the potential educativeness of the post, and which are independent from the semantics of the post (see. Table 2). These are: the Length, Position in thread and Sentiment value.

5. Conclusion

In this paper, we have addressed the task of automatically identifying information potentially useful for the preparation of the educational materials (in particular learning objects learning objects) within technical text forums.

We have overviewed non-semantic text features that indicate suitability for creating LO (relating to chosen topic and containing detailed information), as well as social features from a text forum. The quality of the features was analyzed. Also, linear and non-linear models were constructed. These models were better than baseline selection methods even for forums with small samples on chosen topics and forums with a lot of off-topic text in dataset.

References

1. *Almahy I., Salim N.* (2013), Web Discussion Summarization: Study Review. Proc. of the First International Conference on Advanced Data and Information Engineering (DaEng-2013). Ed. Herawan T. et al. Springer Verlag, 2013. Pp. 649–658.
2. *Banea C., Mihalcea R., Wiebe J.* (2014), Sense-level subjectivity in a multilingual setting. Computer Speech and Language, Vol. 28, pp. 7–19.
3. *Biyani* et al. (2012) Prakhar Biyani, Sumit Bhatia, Cornelia Caragea, Prasenjit Mitra. Thread specific features are helpful for identifying subjectivity orientation of online forum threads. Proc. of the COLING 2012, 24th International Conference on Computational Linguistics, Conference: Technical Papers, 8–15 December 2012, Mumbai, India, pp. 295–310.
4. *Biyani* et al. (2014), Biyani P., Bhati S., Caragea C., Mitra P. Using non-lexical features for identifying factual and opinionative threads in online forums. Knowledge-Based Systems, V. 69, October 2014, pp. 170–178
5. *Boyle, T.* (2003), Developing and delivering learning objects from a practitioner's point of view LTSN, available at: Generic Centre www.heacademy.ac.uk/resouces
6. *Cao* et al. (2011) Yunbo Cao, Wen-Yun Yang, Chin-Yew Lin, Yong Yu. A structural support vector method for extracting contexts and answers of questions from online forums. Information Processing and Management, Vol. 47, pp. 886–898
7. *Carbonaro A.* (2010), WordNet-based Summarization to Enhance Learning Interaction Tutoring. Peer Reviewed Papers. Vol. 6, n. 2, May 2010.
8. *Dringus, Ellis* (2005) Laurie P. Dringus, Timothy Ellis. Using data mining as a strategy for assessing asynchronous discussion forums // Computers & Education, Vol. 45, pp. 141–160
9. *Ferrandez* et al. (2009) Sergio Ferrandez, Antonio Toral, Oscar Ferrandez, Antonio Ferrandez, Rafael Munoz. Exploiting Wikipedia and EuroWordNet to solve Cross-Lingual Question Answering. Information Sciences, Vol. 179, pp. 3473–3488
10. *Freeman L. C.* (1978). Centrality in social networks: Conceptual clarification. Social Networks 1, pp. 215–239.
11. *gbm package* (2014), gbm: Generalized Boosted Regression Models. Available at: http://cran.r-project.org/web/packages/gbm/index.html

12. *Gelman A., Hill J.* (2006), Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press, 2006. ISBN 978-0-521-68689-1
13. *Gradient* boosting machines, a tutorial. Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3885826/
14. *Griffiths J., Stubbs G., Watkins M.* (2007), From course notes to granules: A guide to deriving Learning Object components. Computers in Human, Vol. 23, pp. 2696–2720
15. *Hassan S., Mihalcea R.* (2009), Learning to Identify Educational Materials. International Conference RANLP 2009—Borovets, Bulgaria, pp. 123–127
16. *Hogenboom A.* et al. (2014), Hogenboom A., Heerschop B., Frasincar F., Kaymak U., de Jong F. Multi-lingual support for lexicon-based sentiment analysis guided by semantics. Decision Support Systems, Vol. 62, pp. 43–53.
17. *IEEE* (2002), IEEE 1484.12.1–2002, 15 July 2002, Draft Standard for Learning Object Metadata, IEEE Learning Technology Standards Committee (LTSC)
18. *Järvelin K., Kekäläinen J.* (2002), Cumulated GainBased Evaluation of IR Techniques, ACM Transactions on Information Systems, Vol. 20, No. 4, pp. 422–446.
19. *Kolomiyets O., Moens M.-F.* (2011), A survey on question answering technology from an information retrieval perspective. Information Sciences, Vol. 181, pp. 5412–5434.
20. *Li Y., Liao T., Lai Ch.* (2012), A social recommender mechanism for improving knowledge sharing in online forums. Information Processing and Management, Vol. 48, pp. 978–994
21. *Owczarzak Dang* (2011) Karolina Owczarzak and Hoa Trang Dang. Who wrote What Where: Analyzing the content of human and automatic summaries. Proc. of the Workshop on Automatic Summarization for Different Genres, Media, and Languages, Portland, Oregon, June 23, 2011. Pp. 25–32.
22. *Ren* et al. (2011), Ren Zh., Ma J., Wang Sh. and Liu Y. Summarizing Web Forum Threads based on a Latent Topic Propagation Process. CIKM'11, October 24–28, 2011, Glasgow, Scotland, UK.
23. *Romero C., Ventura S.* (2007), Educational data mining: A survey from 1995 to 2005. Expert Systems with Applications, Vol. 33, pp. 135–146
24. *Romero* et al. (2013) Cristóbal Romero, Manuel-Ignacio López, Jose-María Luna, Sebastián Ventura. Predicting students' final performance from participation in on-line discussion forums. Computers & Education, Vol. 68, pp. 458–472
25. *Said D., Wanas N.* (2011) Clustering posts in online discussion forum threads. International Journal of Computer Science & Information Technology (IJCSIT), Vol. 3, No 2
26. *Smine* et al. (2013) B. Smine, R. Faiz and J-P. Desclés. Relevant learning objects extraction based on semantic annotation. Int. J. Metadata, Semantics and Ontologies, Vol. 8, No. 1, pp. 13–27
27. *Spärck Jones K.* (2007), Automatic summarising: the state of the art. Information Processing and Management, Special Issue on Automatic Summarising, 2007.
28. *Steinberger R.* (2011), A survey of methods to ease the development of highly multilingual text mining. applications. Language Recource and Evaluation, pp. 1–22
29. *Tore Opsahl* (2014). Node Centrality in Weighted Networks. Available at: http://toreopsahl.com/tnet/weighted-networks/node-centrality/

30. *Wang U., Cardie C.* (2011), Summarizing Decisions in Spoken Meetings. Proc. of the Workshop on Automatic Summarization for Different Genres, Media, and Languages, pp. 16–24, Portland, Oregon, June 23, c.
31. *White D. R., Borgatti S. P.* (1994) Betweenness centrality measures for directed graphs/ Social Networks, Vol. 16, pp. 335–346
32. *Wiley D. A., ed.* (2001), Connecting learning objects to instructional design theory: definition, a metaphor, and taxonomy. New York: Agency for Instructional Technology and the Association for Educational Communications and Technology, 2001
33. *Zhou L., Hovy E.* (2005), Digesting virtual "geek" culture: The summarization of technical internet relay chats. Proc. of the 43rd Annual Meeting on Association for Computational Linguistics, ACL '05, pp. 298–305, Stroudsburg, PA, USA.

ЧТО ТАКОЕ *ОРЕХИ*?[1]

Иомдин Б. Л. (iomdin@ruslang.ru)

Институт русского языка имени В. В. Виноградова РАН,
НИУ «Высшая школа экономики», Москва, Россия

В докладе описывается небольшое, но интересное семантическое поле, включающее слово *орех* и названия орехов в русском языке. Существующие описания толковых словарей несистемны и плохо интерпретируются — в первую очередь потому, что в них смешано бытовое и терминологическое словоупотребление, которые в описываемом материале весьма существенно различаются (так, в ботанике орехом обычно считается каштан, но не считаются кедровый орех, арахис, миндаль, фисташка; примерно аналогичная картина с английским термином nut и латинским nux). При этом словоупотребление в ряде областей (в частности, в сфере кулинарии, торговли, медицины), как и в быту, отличается от принятого в ботанике. Был проведен опрос более тысячи респондентов-носителей русского языка и исследованы тексты разных периодов и жанров. В результате были очерчены границы естественного класса природных объектов, называемых *орехами* в русском языке, который по своим свойствам не совпадает ни с одним из типов естественных языковых классов, описанных А. Вежбицкой. С учетом проведенного анализа предложено лексикографическое описание рассматриваемых лексических единиц.

Ключевые слова: семантика, лексикография, полисемия, эволюция значения, опросы, терминология, классификация, бытовое словоупотребление, наивная картина мира, орехи

NUTS: WHAT ARE THEY?

Iomdin B. L. (iomdin@ruslang.ru)

V. V. Vinogradov Russian Language Institute of the Russian Academy of Sciences, National Research University “Higher School of Economics”, Moscow, Russia

[1] Работа выполнена при частичной финансовой поддержке Программы фундаментальных исследований Президиума РАН «Историческая память и российская идентичность», гранта РГНФ №13-04-00307а и гранта НШ-3899.2014.6 для поддержки научных исследований, проводимых ведущими научными школами РФ. Автор также выражает глубокую признательность д. б. н. А. А. Оскольскому за внимательное прочтение работы и ценные замечания и идеи и благодарит А. Ч. Пиперски и анонимных рецензентов работы за замечания и предложения.

When describing words which denote real life objects, dictionaries tend to use scientific terms and classifications, even when dealing with natural language. This approach may lead to misunderstanding, especially in cases when scientific classification (e. g. in biology) differs from what is found in natural language data. One of such cases is discussed here, namely the small but rather interesting class of nuts (Russian *orexi*). In the botanic world view nuts usually include hazelnuts and chestnuts, but do not include almonds (which are considered stone fruits), pine nuts (seeds), peanuts (legumes), pistachio (kernels), etc. The Russian *orex*, English *nut*, Latin *nux* exhibit similar behaviour here. Explanatory dictionaries of Russian more or less follow the botanical definitions, even though in many fields (such as cooking, food industry, medicine, etc.) nuts are classified differently. In order to establish the boundaries of nuts in Russian, more than 1,000 native speakers were questioned and multiple texts of different periods were studied. The result is a peculiar class which could not be identified with any of the natural language supercategories described by Anna Wierzbicka. A new lexicographic description is proposed for some words included into this class.

Key words: semantics, lexicography, polysemy, meaning evolution, surveys, terminology, classificaton, everyday word usage, naïve world view, nuts

В работе [Иомдин 2012], посвященной проблеме несоответствия терминологического и бытового словоупотребления, говорилось о коммуникативных трудностях, возникающих, когда один из партнеров в диалоге использует слово в «неправильном» (общеязыковом), а второй — в «правильном» (профессиональном) значении: «Так, врач и пациент могут по-разному понимать боли в предплечье или аллергию на орехи. Представляется, что и такого рода различия необходимо подробно исследовать и предлагать для них системное лексикографическое описание» [Иомдин 2012: 249]. Настоящая статья посвящена попытке такого описания одного небольшого, но интересного семантического поля — названий орехов в русском языке.

1. *Орехи* в толковых словарях

В БТС, толковом словаре русского языка с одним из наиболее полных словников (не считая, разумеется, пока не законченного БАС 2), слово *орех* встречается всего несколько десятков раз. Сама эта вокабула толкуется следующим образом (опускаем для краткости речения и примеры):

ОРЕХ

1. Плод некоторых растений со съедобным ядром в твёрдой скорлупе, оболочке. // Об очищенных от скорлупы ядрах такого плода. // *Ботан.* Нераскрывающийся плод некоторых растений с твёрдым деревянистым околоплодником (плод дуба, каштана и др.).

2. *только ед.* Дерево или кустарник, дающий такие плоды. // Древесина этого дерева (употребляется в столярном деле из-за её прочности и красоты).

В качестве genus proximum *орех* используется в толковании единственного слова:

ФУНДУК
1. Кустарник или дерево сем. лещиновых; лесной орех.
2. *собир*. Орехи — плоды этого кустарника.

Слова *орех, орешек* или *ореховый* так или иначе упоминаются в толкованиях названий некоторых других природных объектов:

АРАХИС
Травянистое масличное растение сем. бобовых с плодами, содержащими маслянистые вещества; съедобные плоды этого растения; земляной или китайский орех.

ГРЕЦКИЙ ◊ *Грецкий орех*. Южное ореховое дерево; крупный плод этого дерева с очень твёрдой скорлупой.

КАШТАН
Дерево сем. буковых, дающее плоды в виде крупного ореха светло-коричневой окраски; плод такого дерева.

КЕДР
Распространённое неправильное название некоторых видов сосны, дающих съедобные семена — орешки.

КОКОС
1. *Разг*. =Кокосовая пальма.
2. Плод этой пальмы. ◊ *Кокосовый орех* (плод этой пальмы).

МИНДАЛЬ
Южное дерево сем. розоцветных, с розовыми цветками и плодами в виде орехов овальной формы; плод такого дерева.

МУСКАТНИК
Вечнозелёное тропическое дерево с мясистыми жёлтыми плодами, содержащими семя — мускатный орех.

ОРЕШНИК
1. Кустарниковое растение со съедобными плодами — орехами; лещина.
2. Заросли ореховых кустов.

ПЕКАН
Южное дерево сем. ореховых с маслянистыми питательными плодами и ценной древесиной; плод этого дерева.

ФИСТАШКА

Южное дерево или кустарник с перистыми листьями, дающее плоды в виде небольшого ореха; съедобный плод этого дерева.

ЧИЛИМ

Однолетнее травянистое растение сем. рогульниковых, плод которого используют в пищу и на корм скоту; водяной орех (произрастает в Евразии и в Африке).

Слово *орех* также приводится в составе некоторых словосочетаний, оставленных без толкования:

В составе народных названий животных и растений: *чёртовы орехи*.

В составе некоторых ботанических и зоологических названий: *китайский орех*.

Маньчжурский орех — родной брат грецкого ореха.

Наконец, слово *орех* (без уточняющих прилагательных) встречается в толкованиях некоторых кондитерских изделий и т. п.: *грильяж, марципан, нуга, пломбир, пралине, рахат-лукум, халва, чурчхела, шербет*, а также в следующих толкованиях: *щелкунчик* — приспособление для раскалывания (щёлканья) орехов в виде игрушечного человечка с огромным ртом; *ядреный* — имеющий полное, созревшее ядро (об орехе); *ядро* — внутренняя часть плода (обычно ореха), заключённая в твёрдую оболочку.

Легко видеть, что приведенные описания несистемны и трудно интерпретируемы. Так, непонятно, какие именно «некоторые растения» имеются в виду в толкованиях разных оттенков первого значения слова *орех* («плод некоторых растений» vs. «нераскрывающийся плод некоторых растений») и как отличить одни от других. Являются ли орехами «плоды в виде орехов» — каштан, миндаль, фисташка? Если да, то в каком из указанных значений? Являются ли орехами арахис, грецкий орех, кокосовый орех, мускатный орех, пекан, чилим, в словарных описаниях которых эта информация явно не указана? В каком значении употреблено слово *орех* в таких толкованиях, как «*нуга* — кондитерское изделие, сладкая вязкая масса с орехами», и соответственно как определить, какие именно орехи входят в состав нуги?

Подобная же ситуация наблюдается и в других словарях, причем не только русского, но и, в частности, английского языка.

2. *Орехи* как термин

Одна из причин такой несистемности, как уже было сказано, — смешение общеязыкового и терминологического (в данном случае ботанического) словоупотребления. В специальных источниках по ботанике, действительно, орех чаще определяют как «ценокарпный сухой односемянный невскрывающийся плод с сильно одревесневшим околоплодником» [БиоЭС]. Этому описанию

из приведенных природных объектов соответствуют только лесной орех (лещина, фундук) и каштан[2] — чего, как кажется, нельзя определить из описаний БТС. В статье ОРЕХ в БиоЭС специально указывается: «т. н. кокосовый орех — сухая костянка, а грецкий «орех» и «орех» миндаля — косточки сухих костянок». Согласно ботанической литературе, арахис — бобы; кокос и фисташка — костянки; грецкий орех и пекан (кария) — ложные костянки; кедровый орех, мускатный орех и кешью (акажу) — семена.

Впрочем, как справедливо говорится в [Rosengarten 2004], “few botanical terms are used more loosely than the word *nut*”. В специальной ботанической литературе описывается множество переходных случаев употребления терминов *орех* и *орешек*; ср. «Трудности наименования многих видов плодов связаны с недостаточной их изученностью, а также с отсутствием общепринятых критериев для разграничения близких карпологических видов, таких, например, как семянка и орех» [Левина 1987]. Подробный обзор сложнейшей истории употребления терминов nux (лат.) и nut (англ.) в ботанике см. в работе [Spjut 1994], где предложено вообще отказаться от этих терминов: “decided to not use the term “nut” because it has acquired various meanings over time as evident in popular books about “nuts” (e. g., Duke (1989) as well as technical manuals (e. g., Johnson 1931) and in scientific journals (e. g., Judd 1985), which can include seeds as well as fruits (Johnson 1931). In view of the historical confusion over the meanings given to the term nut, especially when botanists continually try find some way to bring them altogether (e. g., Johnson 1931; Judd 1985), it, therefore, seems best to leave the term nut and its varied meanings in the layperson’s realm”.

В терминологических контекстах, не связанных напрямую с ботаникой, употребление рассматриваемых слов тоже различается[3]. Так, определять, что является, а что не является орехами, критически важно, в частности, при наличии аллергии. В мире регистрируется все больше и больше случаев аллергической реакции на орехи (в том числе с летальным исходом); так, по данным телефонных опросов в США, с 1997 по 2008 год количество детей с «аллергией на орехи»

---

[2] По сообщению А. А. Оскольского, «хотя каштан в ботанической литературе обычно рассматривается как орех, он не вполне подходит под это определение, так как у него нет сильно одревеснев шего околоплодника. Если признавать его околоплодник сильно одревеснев шим, то к орехам следует отнести и плоды дуба». Плоды дуба обычно называют желудями и относят к ореховидным плодам [Левина 1987], но иногда и к орехам: «В отличие от других плодов-орехов, например от плода лещины, относимого к ореху многими авторами, и плода дуба, определяемому как орех З. Т. Артюшенко и И. Н. Коноваловым (1951), Ю. Л. Меницким (1984) и Артюшенко и Ал. А. Федоровым (1986), у плода осок экзокарпий не одревеснев ает» [Егорова 1999]; ср. также приведенное выше описание БТС. К семейству буковых, помимо каштана и дуба, относится и бук, чьи плоды чаще называются *орешками*.

[3] Ср. характерную цитату из научного текста: «С ботанической точки зрения словосочетание «орехи кола» некорректно, поскольку в ботанике термин «орех» используется для обозначения одного из типов плода, тогда как «орехами кола» называют семена этого растения. Тем не менее это выражение широко употребляется в повседневной жизни и в коммерции, а потому в данной работе мы будем придерживаться традиционного словоупотребления» [Мищенко, Оскольский 2015].

увеличилось в 3,5 раза и достигло более 6 млн человек [Sicherer et al 2010][4]. Однако остается не вполне ясным, какие именно орехи имеются в виду. В большинстве англоязычных источников идет речь про аллергию на peanuts (арахис) и tree nuts (древесные орехи), с различными уточнениями: “Tree nuts include, but are not limited to, walnut, almond, hazelnut, cashew, pistachio, and Brazil nuts. These are not to be confused or grouped together with peanut, which is a legume, or seeds, such as sunflower or sesame” [сайт FoodAllergy.org]; “Coconut is not a botanical nut; it is classified as a fruit, even though the Food and Drug Administration recognizes coconut as a tree nut. <…> The following are not nuts: nutmeg, water chestnuts and butternut squash” [сайт American College of Allergy, Asthma and Immunology]. Аллергологи подчеркивают, что размытость понятия «nuts» вызывает трудности у пациентов: “The meaning of the terms nuts, seeds and legumes is confusing, particularly for allergic patients (or their parents) trying to decide what foods to avoid” [сайт Allergy.Org.au]; ср. также [Alasavar and Shahidi 2008]. В русскоязычных источниках из этой сферы *орехи* также обычно понимаются не в ботаническом смысле, ср. «Орехи: фундук, бразильский орех, кешью, пекан, фисташки, миндаль, кокос, кедровый орех, грецкий орех» [Лусс 2005]; «В «большую восьмерку» продуктов, обладающих наибольшей аллергенностью, входят <…> орехи (лесные орехи, миндаль, грецкие и др.)» [Боровик и др. 2004].

Казалось бы, если большинство орехов «с точки зрения ботаники являются ненастоящими орехами, или неправильно называемыми орехами» [Википедия], вряд ли оправдано их выделение в отдельную группу по критерию общих свойств как биологических объектов (способность вызывать аллергическую реакцию). Однако интересно, что наличие у пациента аллергии на один из видов орехов с высокой вероятностью означает возможность аллергической реакции и на другие, что связано с так называемой перекрестной реакцией (cross-reactivity) и с наличием в разных орехах общих аллергенов, в том числе липидпереносящих белков (lipid transfer proteins) и запасных белков (storage proteins) [Asero 1999, Goetz et al. 2005]. Проще говоря, природные объекты, которые язык объединяет в единую категорию, обладают и объективным общим свойством — высокой концентрацией белков (что и вызывает аллергические реакции). Не исключено, однако, что на сами исследования перекрестной аллергии на орехи повлияло именно выделение орехов в один класс в естественном языке и что класс биологических объектов, вызывающих перекрестные аллергии, на самом деле больше (серьезное исследование этой проблемы выходит за рамки чисто лингвистической проблематики, но представляется весьма важным для медицины).

Что касается официальной номенклатуры, в частности, в торговой сфере, то там словоупотребление также не соответствует принятому в ботанике. Ср. группы товаров, выделяемые в «Общероссийском классификаторе продукции по видам экономической деятельности», введенном Федеральным агентством по техническому регулированию и метрологии Министерства экономического развития РФ 01.02.2014:

---

[4] В нашем опросе (см. ниже) 14% респонендентов указали, что у них или их близких есть аллергия на орехи.

01.11.8 Бобы соевые, орехи земляные, семена хлопка
01.11.82 Арахис (орех земляной) нелущеный
01.25.3 Орехи, кроме лесных съедобных орехов, земляных орехов и кокосовых орехов
01.25.31 Миндаль
01.25.32 Каштаны
01.25.33 Фундук
01.25.34 Фисташки
01.25.35 Орехи грецкие
01.26 Плоды масличных культур
01.26.20 Орехи кокосовые

Таким образом, особое внимание именно к ботанической терминологии, свойственное толковым словарям, кажется неоправданным, а описание *орехов* должно строиться на данных языка — в частности, на опросах носителей.

3. *Орехи* по данным носителей

В августе-декабре 2014 г. автор провел интернет-опрос с целью уточнить состав класса *орехи* в современном русском языке. Респонденты заполняли анкету на своем родном языке. Всего были получены данные по 28 языкам; о результатах исследования нерусскоязычных респондентов будет сообщено позднее. 1150 респондентов использовали русский язык. Те, кто заполнял анкету на русском языке, указали ок. 300 различных населенных пунктов России и других стран как место постоянного проживания, почти половина участников (48%) из Москвы; средний возраст респондентов — 30 лет.

В первой части опроса требовалось ответить на следующий вопрос: «Какие орехи вы знаете? Перечислите через запятую. Не думайте долго, напишите те, что приходят в голову за 2 минуты». В ответах хорошо выделяется верхняя группа орехов, которые назвали более 100 респондентов (более 10%):

Таблица 1. Какие орехи вы знаете?

фундук / лесной орех / лещина	1067	92,46%
грецкий	1050	90,99%
кешью	942	81,63%
арахис	882	76,43%
миндаль	850	73,66%
кедровый	705	61,09%
фисташка	465	40,29%
бразильский	428	37,09%
кокос	277	24,00%
пекан	257	22,27%
макадамия	153	13,26%

Отметим, что оказавшийся на третьем месте *кешью* отсутствует в рассмотренных нами толковых словарях (МАС, БТС, СЕф, СШ), кроме ТСИ.

Во второй части опроса требовалось опознать каждое изображение из двенадцати приведенных ниже[5].

Таблица 2. Что изображено на картинке? Количество опознавших

11	кокос	1134	98,35%
8	грецкий	1124	97,48%
1	миндаль	1089	94,45%
6	фисташки	1078	93,50%
12	фундук/лесной орех/лещина	1070	92,80%
9	арахис	1006	87,25%
5	кешью	1002	86,90%
4	кедровый	988	85,69%
10	каштан	961	83,35%
7	макадамия	502	43,54%
2	пекан	361	31,31%
3	бразильский	131	11,36%

Наконец, в третьей части опроса респондентам давались названия орехов, приведенные выше, и требовалось ответить на вопрос «Что из этого, по-вашему, не орехи?».

Таблица 3. Что из этого не орехи?

каштан	592	51,39%
кокос	435	37,76%
арахис	265	23,00%
кедровый	140	12,15%
пекан	118	10,24%
макадамия	115	9,98%
кешью	56	4,86%
фисташка	55	4,77%
бразильский	50	4,34%
миндаль	46	3,99%
грецкий	12	1,04%
фундук / лесной орех / лещина	4	0,35%

[5] В исследовании [Ferdman, Church 2004], проведенном в США, детям (аллергикам и здоровым) предъявлялись непосредственно орехи с целью их опознания; лучше всего определили арахис в скорлупе (89%) и без скорлупы (52%), фисташки определили лишь 32%, а бразильские орехи не определили вовсе.

4. *Орехи* как естественный класс

В своей известной работе [Wierzbicka 1984] Анна Вежбицка выделяет пять типов «суперкатегорий», представленных в естественном языке: I таксономические (*птица, цветок, дерево*), II функциональные (*игрушки, оружие*), III коллективные singularia tantum (*мебель*), IV коллективные pluralia tantum (*остатки, покупки*) и V вещества и псевдоисчисляемые (*лекарства, овощи*). Как кажется, *орехи* (во всяком случае, в русском языке) нельзя однозначно отнести ни к одному из перечисленных типов.

На первый взгляд, *орехи* близки к таксономическим категориям I (таким, как *цветок* или *дерево*). В самом деле, большинство говорящих при необходимости определить, что такое *фундук*, по-видимому, скажет что-то вроде «это такой орех». Однако таксономические категории, как указывает Вежбицка, должны иметь явные перцептивные характеристики, которые в случае орехов определить не так легко: можно нарисовать «птицу вообще» или «дерево вообще», но труднее нарисовать «орех вообще» (хотя и проще, чем «игрушку вообще» или «оружие вообще»; ср. [Wierzbicka 1984: 317]). Кроме того, граница между орехами и не-орехами более размыта, чем граница между птицами и не-птицами: ср., напр., *арахис*, который в нашем опросе 76% респондентов указали в числе орехов, первыми приходящих на ум, а 23% респондентов назвали «не-орехами».

С другой стороны, *орехи* не являются и функциональной категорией II, куда относятся *игрушки* и *оружие* — прежде всего потому, что орехи не артефакты. Отчасти *орехи* обладают свойством класса III, которое Вежбицка называет contiguity (смежность): “kitchenware includes things of different kinds used jointly in the kitchen; bedlinen, things of different kinds put jointly on a bed” [Wierzbicka 1984: 321]. В самом деле, разумно считать, что миндаль, фундук и кешью называют орехами не потому, что они внешне похожи (как похожи друг на друга птицы или деревья), а в том числе и потому, что мы их часто видим и используем в одних и тех же ситуациях (скажем, на тарелке, поданной среди легких закусок, или на одной полке в магазине); по этой же причине мы обычно не называем орехами каштаны (в первом вопросе анкеты каштан привели лишь 5% респондентов), и внешне, и по своим ботаническим свойствам похожие на тот же фундук. Но при этом *орехи*, по-видимому, не являются гетерогенным классом, как посуда или мебель: так, фраза *Мне надо купить орехов* не обязательно означает, что мне надо купить орехов разных типов (ср. “A sentence such as, I must buy some cutlery/furniture, would be interpreted as referring to more than one kind of thing. By contrast, a sentence such as, I must buy some flowers, would not be interpreted as implying more than one kind of thing” [Wierzbicka 1984: 320]). Не относятся *орехи* и к классу IV (*остатки, покупки* и т. п.)[6]. Наконец,

---

[6] Отметим, впрочем, выражение nuts-and-bolts, которое Вежбицка приводит в качестве примера как раз на этот класс, указывая в скобках значение «party snacks». Это значение данного выражения не дает ни один из рассмотренных нами английских словарей (все они приводят лишь значение ‘основные детали, основные элементы’), однако оно явно построено на омонимии nut ‘гайка’ и nut ‘орех’ и встречается, в частности, в рецептах на сайте Peanut Company of Australia (что неудивительно, принимая во внимание тот факт, что Анна Вежбицка живет и работает в Австралии); см. выше наше соображение об орехах как отдельном типе легких закусок.

орехи имеют общие свойства и со словами класса V (*овощи*), ср. часть признаков этого класса, указанных Вежбицкой: “things of different kinds which are used for the same kind of purpose and which people have come to have in the same kind of way” [Wierzbicka 1984: 327]. Однако они не обладают свойством неисчисляемости и могут называть не типы объектов, а сами объекты. Ср. приводимую Вежбицкой фразу I like only three vegetables: spinach, broccoli and celery, но странную ?*Я люблю только три ореха: фундук, миндаль и арахис*; напротив, если фраза I had three vegetables for dinner, по мнению Вежбицкой, не может описывать ситуацию ‘Я съел одну свеклу и две морковки’, то фраза *Я съел три ореха*, на наш взгляд, вполне может описывать ситуацию ‘Я съел два грецких ореха и один лесной’.

Таким образом, *орехи*, с одной стороны, обладают частью свойств таксономических категорий, что позволяет включать это слово в толкования других слов, таких как *фундук*, *миндаль* и др.; с другой стороны, частью свойств коллективных категорий, что требует включения в толкование слова *орехи* указание на их частое применение в схожих целях и совместно; и, с третьей стороны, частью свойств категорий типа *овощи*, что требует указание на их происхождение и использование (ср. толкование слова vegetable, приводимое А. Вежбицкой: “a thing of any kind that people cause to grow out of the ground for people to cook for food” [Wierzbicka 1984: 323]).

По-видимому, дело в том, что *орехи* — достаточно новый класс, формирующийся на наших глазах. В русском языке он возникает в результате расширения значения слова *орех*, первоначально относящегося только к лесному ореху (и другим близким видам рода Corylus[7]; ср. подробно в Cooper 1999: 301), затем распространившегося и на грецкий орех (и другие близкие виды рода Juglans[8]), семена некоторых видов сосны, Pinus [СРЯ XI–XVII] и постепенно охватывающего и другие виды биологических объектов, по мере их перехода из разряда экзотики в разряд привычных продуктов питания. Ср. эволюцию слова *ягода*, в древнерусском языке, по-видимому, означавшего только ягоду винограда, а во многих славянских языках и сейчас означающего только ягоду[9] земляники (клубники) [Иванов 1989]. Очевидно, что в разных языках такие естественные классы устроены по-разному; ср., напр. “We also find conceptual gaps in the case of a lack of Spanish or French term equivalent to the English nut, which includes walnuts, peanuts, almonds, etc., because there is not a natural category for such a thing in either language” [Goded Rambaud 2012: 141]; ср. подробнее также о естественных классах в разных языках в [Iomdin et al. 2011].

Постепенность этого процесса можно заметить и по снижению встречаемости в корпусах текстов сочинительных цепочек вида *орехи и миндаль*, *орехи*

---

[7] Характерно, что *лесной орех* назывался иногда *простой орех* или *обыкновенный орех*.

[8] Это, в частности, отразилось в появлении у рода растений Juglans (не вида плодов — напомним, что плоды грецкого ореха ботаники называют ложными костянками!) ботанического наименования *орех* и названия семейства *ореховые*, что привело к дополнительной путанице.

[9] Нельзя не отметить, что с ботанической точки зрения плод земляники — не ягода, а многоорешек.

и фисташки и т. п.: чем старше тексты, тем легче обнаруживаются такие примеры. Сплошный просмотр всех употреблений слова *орех* в основном подкорпусе НКРЯ позволил обнаружить 12 таких примеров в текстах за период с 1869 по 1939 год и только 3 — за период с 1940 по 2011 год (притом что общий объем текстов в первой группе почти в 2 раза меньше, чем во второй). Ср. цитаты из НКРЯ:

> *Кроме всех упомянутых кушаньев, еще вредны детям все невареные овощи, конфекты, сырой сахар, а особливо крашеные сахарные товары,* ***миндаль, орехи,*** *изюм и другие такие лакомства* [Н. И. Новиков. О воспитании и наставлении детей (1783)]

> *Народ неистово накидывается на* ***миндаль, пастилу и орехи****, хватает горстями просыпанный чай, набивает карманы разными продуктами* [В. В. Крестовский. Панургово стадо (1869)]

> *Сидели, слушали и в то же время щелкали* ***орехи, фисташки, миндаль*** [Н. Г. Гарин-Михайловский. Гимназисты (1895)]

> *Мне предложено было заботиться о сладком, т. е. изюме,* ***миндале, орехах*** *и т. д.* [В. В. Верещагин. Дунай. 1877 (1899)]

Одновременно сокращается число примеров, в которых *орех* относится именно к лесному ореху; чем новее тексты, тем чаще слово орех используется гиперонимически, и *миндаль*, *фисташки* и др. явно включаются в состав *орехов*:

> *Сказала, что из всех* ***орехов*** *больше всего любит* ***фисташки****, от шоколада отказалась* [М. Голованивская. Противоречие по сути (2000)]

> *Фрукты — все, кроме слив, абрикосов, черешни, бананов и смородины.* ***Орехи*** *— несоленые* ***арахис, миндаль и фундук****. Хороши сладости, варенье, душистые травы* [«Семейный доктор», 2002.05.15]

> *Этот микроэлемент есть в молоке, твороге, морской рыбе,* ***орехах (арахисе, миндале)*** *и сухофруктах* [«Семейный доктор», 2002.07.15]

Особо отметим, что это не касается *каштанов*, которые и в современных текстах употребляются в сочинительных цепочках с *орехами*:

> *Заполненный взрослыми и детьми перекресток пахнет жареными* ***каштанами****, вином* ***и орехами*** [Андрей Дмитриев. Закрытая книга (1999)]

> *Питались* ***каштанами, орехами****. Заслышав кабанов, отсиживались на деревьях* [Александр Иличевский. Матисс // «Новый Мир», 2007]

5. *Орехи*: лексикографическое описание

В заключение предлагаем примерное лексикографическое описание слова *орехи* и некоторых названий орехов в современном русском языке в формате Активного словаря русского языка [Апресян и др. 2014], в сокращенном виде (без примеров, иллюстраций, сведений о сочетаемости и др.).

ОРЕХ

орех 1.1

ЗНАЧЕНИЕ.

‘Сухой и твердый питательный плод, содержащий большое количество белков и жиров, который можно есть сырым и который обычно можно поместить в рот целиком’.

КОММЕНТАРИИ.

1. Орехами обычно называют следующие плоды: фундук (лесной орех, лещина), грецкий орех, кешью, арахис (земляной орех), миндаль (миндальный орех), кедровый орех, фисташка.
2. Разные орехи обычно продают в одном и том же месте.
3. Разные орехи часто подают к столу вместе как вид закуски.

орех 1.2, бот.

ЗНАЧЕНИЕ.

‘Сухой невскрывающийся плод некоторых деревьев или кустарников с одним съедобным ядром в твердой оболочке’.

КОММЕНТАРИИ.

В специальных текстах по ботанике орехами обычно называют плоды следующих деревьев: граб, лещина, каштан, бук.

орех 2.1

ЗНАЧЕНИЕ.

‘Дерево семейства ореховых’.

КОММЕНТАРИИ.

Наиболее известные плоды ореха — грецкие орехи и орехи пекан.

орех 2.2

ЗНАЧЕНИЕ.

‘Древесина ореха 2.1’.

◊ ***кокосовый орех*** ‘плод кокосовой пальмы’; ***мускатный орех*** 1) ‘дерево семейства мускатниковых’; 2) ‘семя этого дерева, обычно использующееся в молотом виде как пряность’.

ФУНДУК

фундук 1

ЗНАЧЕНИЕ.

‘Орех размером с ноготь взрослого человека, растущий на дереве, в твердой оболочке, круглой формы, коричневого цвета, который едят очищенным’.

АНАЛОГ: *лесной орех*.

фундук 2
ЗНАЧЕНИЕ.
‘Дерево семейства березовых, на котором растет фундук 1’.
АНАЛОГ: *лещина*.

КЕШЬЮ
кешью 1
ЗНАЧЕНИЕ.
‘Орех размером с фалангу пальца взрослого человека, растущий на дереве, в ядовитой оболочке, изогнутой формы, желтовато-белого цвета и сладковатого вкуса, который едят очищенным и обжаренным’.
СИНОНИМ: *акажу*.
кешью 2
ЗНАЧЕНИЕ.
‘Тропическое дерево семейства сумаховых, на котором растет кешью 1’.
КОММЕНТАРИИ.
В специальных текстах по ботанике семена кешью, растущие на концах разросшейся грушевидной плодоножки, не называются орехами.
СИНОНИМ: *анакардиум*.

АРАХИС
арахис 1
ЗНАЧЕНИЕ.
‘Орех размером чуть больше ногтя взрослого человека, растущий на траве в бобах, дозревающих в земле, овальной формы, покрытый тонкой красноватой кожицей, нейтрального вкуса, который едят очищенным и обычно обжаренным’.
СИНОНИМ: *земляной орех*.
арахис 2
ЗНАЧЕНИЕ.
‘Южноамериканское однолетнее травянистое растение семейства бобовых, на котором растет арахис 1’.
КОММЕНТАРИИ.
В научных текстах семена арахиса не называются орехами.

МИНДАЛЬ
миндаль 1
ЗНАЧЕНИЕ.
‘Орех размером с фалангу пальца взрослого человека, растущий на дереве, в твердой оболочке, овальный с заостренными краями, темно-бежевого цвета и горьковатого вкуса, который едят очищенным’.
миндаль 2
ЗНАЧЕНИЕ.
‘Южное дерево семейства розовых, на котором растет миндаль 1’.

КОММЕНТАРИИ.

В специальных текстах по ботанике плоды миндаля называются не орехами, а сухими костянками со съедобными косточками (подобно плодам абрикоса, сливы, вишни, черешни, входящих в то же семейство).

ФИСТАШКА

фисташка 1

ЗНАЧЕНИЕ.

'Орех размером с ноготь взрослого человека, растущий на дереве, в твердой оболочке, обычно раскрывающейся при созревании на две створки, овальной формы, светло-зеленого цвета, сладковатого вкуса, который едят очищенным'.

фисташка 2

ЗНАЧЕНИЕ.

'Тропическое дерево семейства сумаховых, на котором растет фисташка 1'.

КОММЕНТАРИИ.

В специальных текстах по ботанике семена фисташки не называются орехами.

Описания могут быть уточнены в зависимости от целей конкретного словаря, однако представляется важным различение общеязыкового и специального словоупотребления. Как кажется, при последовательном применении этого принципа можно обнаружить и многие другие примеры естественноязыковых классов, не совпадающих с традиционно выделяемыми в науке, которые еще ждут своего исследования.

Литература

1. *Апресян и др. 2014* — Апресян В. Ю., Апресян Ю. Д., Бабаева Е. Э., Богуславская О. Ю., Галактионова И. В., Гловинская М. Я., Иомдин Б. Л., Крылова Т. В., Левонтина И. Б., Лопухина А. А., Птенцова А. В., Санников А. В., Урысон Е. В. Активный словарь русского языка. Тт. 1–2: А–Г. Под ред. Ю. Д. Апресяна. М.: «Языки славянских культур», 2014.
2. *БАС 2* — Большой академический словарь русского языка / ИЛИ РАН; Под ред. К. С. Горбачевича. М, СПб.: «Наука», 2004–2014. Тт. 1–23.
3. *БиоЭС* — Биологический энциклопедический словарь. Гл. ред. М. С. Гиляров. М.: Сов. энциклопедия, 1986.
4. *Боровик и др. 2004* — Боровик Т. Э., Лаврова Т. Е., Ревякина В. А., Рославцева Е. А. Современный взгляд на проблему пищевой непереносимости. Вопросы современной педиатрии. Выпуск № 6, том 3, 2004. С. 40–49.
5. *БТС* — Большой толковый словарь русского языка / Сост., гл. ред. С. А. Кузнецов. СПб.: Норинт, 1998.
6. *Егорова 1999* — Егорова Т. В. Осоки (Carex L.) России и сопредельных государств (в пределах бывшего СССР). СПб.: Санкт-Петербургская ГХФА, Сент-Луис: Миссурийский ботанический сад, 1999.

7. *Иванов 1989* — Иванов В. В. Общеславянский лингвистический атлас : материалы и исследования: 1985–1987. М.: Наука, 1989.
8. *Иомдин 2012* — Иомдин Б. Л. О «неправильном» использовании терминов: может ли язык ошибаться? // Смыслы, тексты и другие захватывающие сюжеты: Сб. ст. в честь 80-летия И. А. Мельчука. / Под ред. Ю. Д. Апресяна, И. М. Богуславского, Л. Ваннера, Л. Л. Иомдина, Я. Миличевич, М.-К. Л'Омм, А. Польгера. М.: Языки славянской культуры, 2012. С. 233–251.
9. *Левина 1987* — Левина Р. Е. Морфология и экология плодов. Л.: Наука, 1987.
10. *Лусс 2005* — Лусс Л. В. Пищевая аллергия и пищевая непереносимость: терминология, классификация, проблемы диагностики и терапия. Учебное пособие. М.: Фармарус Принт, 2005.
11. *МАС* — Словарь русского языка: В 4-х т. /АН СССР, Ин-т рус. яз.; Под ред. А. П. Евгеньевой. М.: Русский язык, 1985–1988.
12. *Мищенко Д. Ф., Оскольский А. А.* Кола у народов манде // Бетель, кава, кола, чат. Жевательные стимуляторы в ритуале и мифологии народов мира. Маклаевский сборник. Вып. 5. СПб.: МАЭ РАН, 2015 (в печати).
13. *СЕф* — Ефремова Т. Ф. Большой современный толковый словарь русского языка. В 3-х т. М.: АСТ, Астрель, 2006.
14. *СРЯ XI–XVII* — Словарь русского языка XI–XVII вв. Вып. 13 (Опасъ — Отработыватися). М.: Наука, 1987.
15. *СШ* — Толковый словарь русского языка с включением сведений о происхождении слов / Отв. ред. Н. Ю. Шведова. М.: Азбуковник, 2007.
16. *ТСИ* — Крысин Л. П. Толковый словарь иноязычных слов. М.: Русский язык, 1998.
17. *Alasavar and Shahidi 2008* — Tree Nuts: Composition, Phytochemicals, and Health Effects. Ed. by C. Alasalvar, F. Shahidi. Boca Raton, CRC Press, 2008.
18. *Asero 1999* — Asero R. Detection and clinical characterization of patients with oral allergy syndrome caused by stable allergens in Rosaceae and nuts. Annals of Allergy, Asthma & Immunology, 1999, 83(5), 377–383.
19. *Cooper 1999* — Cooper B. Hazel and some other nut terms in Russian. New Zealand Slavonic Journal. 1999. Pp. 297–318.
20. *Ferdman and Church 2006* — Ferdman R. M., Church J. A. Mixed-up nuts: identification of peanuts and tree nuts by children. Annals of Allergy, Asthma & Immunology, 2006, 97(1), 73–77.
21. *Goetz et al. 2005* — Goetz D. W., Whisman B. A., Goetz A. D. Cross-reactivity among edible nuts: double immunodiffusion, crossed immunoelectrophoresis, and human specific IgE serologic surveys //Annals of Allergy, Asthma & Immunology. 2005. Vol. 95. No. 1. Pp. 45–52.
22. *Goded Rambaud 2012* — Goded Rambaud M. Basic semantics. Universidad Nacional de Educación a Distancia. Madrid, 2012.
23. *Iomdin et al 2011* — Iomdin B., Piperski A., Russo M., Somin A. How different languages categorize everyday items. In: Computational linguistics and intellectual technologies. Papers from the annual international conference "Dialogue" (2011). Moscow: RGGU, 2011, p. 258–268.
24. *Rosengarten 2004* — Rosengarten F. The book of edible nuts. Dover publications, 2004.

25. *Sicherer et al. 2010* — Sicherer S. H., Muñoz-Furlong A., Godbold J. H., Sampson H. A. US prevalence of self-reported peanut, tree nut, and sesame allergy: 11-year follow-up. J Allergy Clin Immunol. 2010 Jun; 125(6):1322–6.
26. *Spjut 1994* — Spjut R. A systematic treatment of fruit types. New York: The New York Botanical Garden, 1994.
27. *Wierzbicka 1984* — Wierzbicka A. Apples are not a "kind of fruit": The semantics of human categorization. American Ethnologist, 11(2), 313–328.

References

1. *Alasavar C., Shahidi F.* (eds.) (2008), Tree nuts: composition, phytochemicals, and health effects, CRC Press, Boca Raton.
2. *Apresyan V. Yu., Apresyan Yu. D., Babaeva E. E., Boguslavskaya O. Yu., Galaktionova I. V., Glovinskaya M. Ya., Iomdin B. L., Krylova T. V., Levontina I. B., Lopukhina A. A., Ptentsova A. V., Sannikov A. V., Uryson E. V.* (2014), Active dictionary of Russian language [Aktivnyy slovar' russkogo yazyka]. Vol. 1–2: A–G. Ed. by Yu. D. Apresyan. Yazyki slavyanskikh kul'tur, Moscow, 2014.
3. *Asero R.* (1999), Detection and clinical characterization of patients with oral allergy syndrome caused by stable allergens in Rosaceae and nuts. Annals of Allergy, Asthma & Immunology, 83(5), pp. 377–383.
4. *Borovik T. E., Lavrova T. E., Revyakina V. A., Roslavtseva E. A.* (2004), The current view on the food intolerance [Sovremennyy vzglyad na problemu pishchevoy neperenosimosti], Issues in Contemporary Pediatrics [Voprosy sovremennoj pediatrii], issue 6, Vol. 3.
5. *Cooper B.* (1999), Hazel and some other nut terms in Russian, New Zealand Slavonic Journal, pp. 297–318.
6. *Efremova T. F.* (2006), Great contemporary explanatory dictionary of Russian language. In 3 volumes [Bol'shoy sovremennyy tolkovyy slovar' russkogo yazyka. V 3 t], AST Astrel', Moscow.
7. *Egorova T. V.* (1999), Sedges (Carex L.) of Russia and adjacent states (within the limits of the former USSR) [Osoki (Carex L.) Rossii i sopredel'nykh gosudarstv (v predelakh byvshego SSSR)], Missouri Botanical Garden Press, St. Louis, MO.
8. *Evgen'eva A.P.* (ed.) (1985–1988), Dictionary of Russian language in 4 volumes [Slovar' russkogo yazyka v 4 tomakh], Russkiy yazyk, Moscow.
9. *Ferdman R. M., Church J. A.* (2006), Mixed-up nuts: identification of peanuts and tree nuts by children, Annals of Allergy, Asthma & Immunology, 97(1), pp. 73–77.
10. *Gilyarov M. S.* (ed.) (1986), Biologicheskiy entsiklopedicheskiy slovar' [Biological encyclopedic dictionary], Sovetskaya entsiklopediya, Moscow.
11. *Goded Rambaud M.* (2012), Basic semantics, Universidad Nacional de Educación a Distancia, Madrid.
12. *Goetz D. W., Whisman B. A., Goetz A. D.* (2005), Cross-reactivity among edible nuts: double immunodiffusion, crossed immunoelectrophoresis, and human specific IgE serologic surveys, Annals of Allergy, Asthma & Immunology, Vol. 95., No. 1, pp. 45–52.

13. *Gorbachevich K. S.* (ed.) (2004–2013), Great academic dictionary of Russian language [Bol'shoy akademicheskiy slovar' russkogo yazyka], Vol. 1–22, Nauka, St. Petersburg.
14. *Iomdin B., Piperski A., Russo M., Somin A.* (2011). How different languages categorize everyday items, Computational linguistics and intellectual technologies, papers from the annual international conference "Dialogue", RGGU, Moscow, pp. 258–268.
15. *Iomdin B. L.* (2012), On the "incorrect" usage of terms: can language be wrong? [O "nepravil'nom" ispol'zovanii terminov: mozhet li yazyk oshibat'sya?], Meanings, texts and other exciting things: a Festschrift to commemorate the 80th anniversary of professor I. A. Mel 'chuk [Smysly, teksty i drugie zakhvatyvayushchie syuzhety: Sb. st. v chest' 80-letiya I. A. Mel'chuka], Yazyki slavyanskoy kul'tury, pp. 233–251.
16. *Ivanov V. V.* (1989), Slavic linguistic atlas: materials and research: 1985–1987 [Obshcheslavyanskiy lingvisticheskiy atlas: materialy i issledovaniya: 1985–1987], Nauka, Moscow.
17. *Krysin L. P.* (1998). Explanatory dictionary of foreign words [Tolkovyy slovar' inoyazychnykh slov], Russkiy yazyk, Moscow.
18. *Kuznetsov S. A.* (ed.) (1998), Great explanatory dictionary of Russian language [Bol'shoy tolkovyy slovar' russkogo yazyka], Norint, St. Petersburg.
19. *Levina R. E.* (1987), Morphology and ecology of fruits [Morfologiya i ekologiya plodov], Nauka, Leningrad.
20. *Luss L. V.* (2005), Food allergy and food intolerance: terminology, classifications, diagnostics and therapy [Pishchevaya allergiya i pishchevaya neperenosimost': terminologiya, klassifikatsiya, problemy diagnostiki i terapiya].
21. *Mishchenko D. F., Oskol'skiy A.A.* (2015), Kola nuts in Mande tribes [Kola u narodov mande], Betel, kava, kola, chat. Chewing stimulants in rites and mythology of world peoples [Betel', kava, kola, chat. Zhevatel'nye stimulyatory v rituale i mifologii narodov mira], Vol. 5, MAE RAN, St. Petersburg (in print).
22. *Rosengarten F.* (2004), The book of edible nuts, Dover publications, New York, NY.
23. *Shmelev D. N.* (ed.) (1987). Dictionary of Russian language of 11th-17th centuries [Slovar' russkogo yazyka XI–XVII vv.], Vol. 13, Nauka, Moscow.
24. *Shvedova N. Yu.* (2007). Explanatory dictionary of Russian with etymological data [Tolkovyy slovar' russkogo yazyka s vklyucheniem svedeniy o proiskhozhdenii slov], Azbukovnik, Moscow.
25. *Sicherer S. H., Muñoz-Furlong A., Godbold J. H., Sampson H. A.* (2010), US prevalence of self-reported peanut, tree nut, and sesame allergy: 11-year follow-up, J Allergy Clin Immunol, 125(6), pp. 1322–1326.
26. *Spjut R.* (1994), A systematic treatment of fruit types. The New York Botanical Garden, New York, NY.
27. *Wierzbicka A.* (1984), Apples are not a "kind of fruit": The semantics of human categorization, American Ethnologist, 11(2), pp. 313–328.

ПРОБЛЕМА НЕДИСКРЕТНОСТИ И СТРУКТУРА УСТНОГО ДИСКУРСА

Кибрик А. А. (aakibrik@gmail.com)

Институт языкознания РАН;
МГУ им. М. В. Ломоносова, Москва, Россия

THE PROBLEM OF NON-DISCRETENESS AND SPOKEN DISCOURSE STRUCTURE[1]

Kibrik A. A. (aakibrik@gmail.com)

Institute of Linguistics RAS;
Lomonosov Moscow State University, Moscow, Russia

Language consists of units of various hierarchical levels, but the boundaries between the units are not always crisp, and non-discrete effect are observed. That applies not only to syntagmatic structure, but also to paradigmatics, diachrony, and even whole languages. Non-discreteness is a common property of language and cognition. In contrast to conventional discrete and continuous structures, I propose another kind of structure that can be called focal. Focal phenomena are simultaneously distinct and related. It is necessary to recognize focal structure as one of the major types of structures typical of natural language. Non-discrete effects can be observed at the level of discourse. Spoken discourse consists of elementary discourse units (EDUs), identifiable with the help of a set of behavioral criteria. Along with prototypical clausal EDUs, there are deviant EDUs of various kinds. Parcellated elaborations constitute an example of a paradigmatic outlier among the EDUs. Non-discrete boundaries between EDUs are an illustration of syntagmatic difficulties in EDU identification. Phonemes, EDUs, and other units are not as crisp and clean as our digital mind would want them to be. In order to address linguistic reality in its actual complexity, we have to recognize that segmentation follows the principles of focal structure, which is the general property of language and cognition.

Key words: discreteness, non-discreteness, theory of language, discourse structure

[1] This study was supported by grant 13-06-00179 from the Russian Foundation for Basic Research.

1. Non-discreteness in language and focal structure

Linguists tend to think about language as a system of discrete, segmental units (phonemes, morphemes, words, sentences...). But this view, in its pure form, does not survive an encounter with reality. For example, phoneticians are well aware of the phenomenon of coarticulation. To take a random example, Engwall (2000) demonstrated in an articulographic study that the pronunciation of Swedish fricative consonants is strongly affected by the surrounding vowels. In particular, the context of labial vowels strongly increases lip protrusion, while the context of the front vowel /ɪ/, compared to back vowels, leads to a more anterior position of the tongue (Engwall 2000: 10). These kinds of facts, common in phonetic syntagmatic structure, indicate that speakers, when pronouncing a phoneme, simultaneously pronounce a neigboring phoneme. Boundaries between segments are not always segmental, and trying to posit boundaries in the signal inevitably means a kind of digitization.

In Kibrik 2012a, 2013 I demonstrate that similar kinds of phenomena occur at various syntagmatic levels of language, including sequences of morphemes, words, phrases, etc. Non-discrete effects occur in paradigmatics as well. For instance, Russian may be claimed to have a marginal phoneme /w/, e.g. in rendering English names such as *William* or English borrowings such as *wow*. In fact, in paradigmatics, and especially in semantic paradigmatics, non-discrete effects have been subject to substantial theoretical consideration, cf. Wittgenstein 1953/2001, Labov 1973, Rosch 1973, Lakoff 1987, Zaliznjak 2006, Janda 2015, among others. Of course, if one turns to the diachronic dimension of language, non-discrete phenomena abound here as well. For example, the English *weed* in the idiomatic expression *widow's weed* 'a widow's mourning clothes' can be historically connected to two Old English sources: *wēod* 'plant' and *wǣd(e)* 'garment' (Hock and Joseph 1996: 237–238). Moving from particular linguistic elements to whole languages, we again encounter non-discrete effects. Cienki (2015) argues that the notion of language in general is a prtototype-like category. Particular human languages resist discrete identification both synchronically (the language/dialect problem) and diachronically. Is there a discrete boundary between Russian and Belorussian, or between Old Russian and Russian? Linguists often underestimate non-discrete effects at the level of whole languages. Questioning the validity and integrity of the notion of Common Nordic, Dahl wittily remarks that authors sometimes seem to assume that the Scandinavians "changed their language all at the same time and in the same fashion, as if conforming to a EU regulation on the length of cucumbers" (2001: 227). Of course, the problem of language boundaries is further affected by language contact, blurring the classical crisp family tree model; to cite just one example, Trudgill (2011: 56–58) demonstrates how Scandinavian languages were affected by Low German. Non-discrete effects are not limited to language only but extend to cognition in general. For example, Alexandrov and Sergienko (2003) suggest that psychophysiological experiments prove the non-disjunctive character of mind and behavior; "continuity is the overarching principle in the organization of living things at various levels" (2003: 105). Van Deemter (2010) provides a book-long account of various vagueness-related effects in language and cognition.

Summarizing what has been said so far, language (as well as cognition in general) simultaneously longs for discrete, segmented structure and tries to avoid it. The

omnipresence of non-discrete effects has not yet led to proper recognition in the mainstream linguistic thinking. In fact, linguists are often bashful about non-discreteness. But non-discreteness is not just a nuisance that can be somehow avoided. Non-discrete effects permeate every single aspect of language, and this problem is in the core of theoretical debates about language. Main reactions to this problem can be generally grouped into two types. First, there is a strong tradition of what can be called "digital" linguistics, ignoring non-discrete phenomena or dismissing them as minor. This tradition is associated with Ferdinand de Saussure's motto that language only consists of identities and differences. This tradition has an appeal of scientific rigor but suffers from strong reductionism. In contrast, there is a tradition of inclusive, or "analog", linguistics. This tradition is more realistic but often boils down to a mere statement of continuous boundaries and countless intermediate/borderline cases. I propose that in the case of language we see a special kind of structure that combines the properties of discrete and non-discrete and can be dubbed **focal** structure. Focal phenomena are simultaneously distinct and related. One should not be forced to choose between discrete and continuous structure as the only two available options. This kind of sharp opposition is sometimes proposed by the advocates of the strictly discrete approach, e.g. by Goddard (2011: 233) in his attempt to defend the discrete character of meaning by dismissing the idea of a continuum or merging.

Focal structure is the hallmark of linguistic and, more generally, cognitive phenomena, in contrast to simpler kinds of matter. Focal structure, as well as two other kinds of structure, are represented on Fig. 1. In fact, focal structure can be viewed as the underlying type of structure, discrete and continuous structures being special cases.

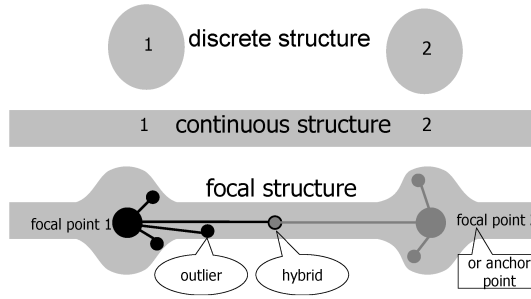


Fig. 1. Various kinds of structures

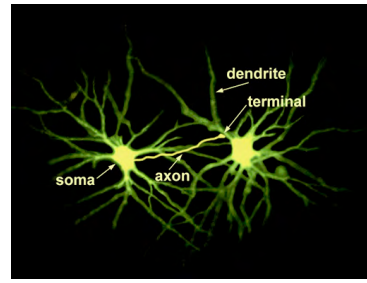


Fig. 2. Neuronal structure with synapses

A possible analogy to focal structure is observed in the neuronal network that serves as the brain substrate of language and cognition, see Fig. 2. This may be more than a mere analogy; the similarity is too obvious to be purely coincidental. At a higher level of brain organization, there is accumulating evidence that neuronal minicolumns may be arranged in two different ways: widely spaced minicolumns, primarily in the left hemisphere, function as discrete units, whereas narrow spacing of minicolumns, mostly in the right hemishpere, is responsible for holistic processing; moreover, the two streams of processing may occur in parallel due to the connection via corpus callosum (see Chance 2014 for a review).

Why are linguists, definitely aware of the non-discrete effects, so much inclined to ignore them? The answer is probably related to the well known Kant's problem. In his "Critique of Pure Reason" Kant suggested that the observer, or cognizer, crucially affects the knowledge of the world: "The schematicism by which our understanding deals with the phenomenal world <...> is a skill so deeply hidden in the human soul that we shall hardly guess the secret trick that Nature here employs." The human analytical mind is digital, and it wants its object of observation to be digital as well. We clearly face here what Dawkins (2011) called "the tyranny of the discontinuous mind". This may be partly because of the scientific tradition based on segmentation and categorization (Aristotelian, "rational", "left-hemispheric", etc.).

2. Segments of talk: Elementary discourse units

What can be done to mend the situation, that is to move towards a more realistic approach to language? We need to develop a more embracing linguistics and cognitive science that address non-discrete phenomena not as exceptions or periphery of language and cognition but rather as their core. In Kibrik 2012a, 2013 I proposed two possible avenues that can help to reach this goal. First, to somewhat shift the primary object of investigation: concentrate on those linguistic phenomena that are less burdened with the tradition of discrete analysis. Second, to entertain new types of models and, possibly, new mathematics, appropriate for the "cognitive matter". In the rest of this paper I make some steps along the first avenue, discussing non-discrete effects in spoken discourse.

In addition to the traditional hierarchical levels of language, including phonetics, moprhology, and syntax, there is a further level of discourse. I have reviewed above the non-discrete effects found at the traditional levels of language. Let us consider the level of discourse, in particular spoken discourse, which is a relatively new object of study in linguistics. I begin with the instances in which spoken discourse displays a kind of segmented structure and proceed with discussing non-discrete and focal effects in the following two sections.

As in other hierarchical levels of language, one can identify discourse segments—intonation units (Chafe 1994) or elementary discourse units (EDUs, Kibrik and Podlesskaya eds. 2009). As many other procedures in the analysis of human behavior, segmentation into EDUs is based on expert assessment and cannot be fully formalized. EDUs are identified by trained experts with the help of a set of behavioral criteria, associated

with the speaker's patterns or vocalization and prosody: pausing, tempo, loudness, intonation, and accent placement. Thus identified EDUs display a remarkable correlation with independently established semantic and syntactic units, that is clauses. See Table 1 for the data from several languages, explored from this perspective.

Table 1. Share of clausal EDUs in various languages

Language	Percentage of clausal EDUs
English (Chafe 1994)	60.0%
Mandarin (Iwasaki and Tao 1993)	39.8%
Sasak (Wouk 2008)	51.7%
Japanese (Matsumoto 2003)	68.0%
Russian (Kibrik and Podlesskaya eds. 2009)	68.6%
Upper Kuskokwim (Kibrik 2012b)	70.8%

Differences across the numerical results for various languages, shown in Table 1, may be partly due to differences between the languages, but also to differences between the explored discourse types, as well as differences between the specific procedures of EDU and clause identification.

Let me provide one short English example consisting of two clausal EDUs and illustrating the basic generalization. (All examples cited in this paper are taken from text SBC032 of the Santa Barbara corpus of spoken American English, see http://www.linguistics.ucsb.edu/research/santa-barbara-corpus. Transcription conventions are the same as we use for Russian discourse, see http://spokencorpora.ru/showtranshelp.py.)

(1) 60.57 45 ••••(3.38) And /then I was /f-fforced \out,
65.84 46 ••(0.07) because I /failed a /promotion to /\commander!

In terms of focal structure introduced above, clausal EDUs in (1) illustrate focal points, that is canonical instances.

3. Parcellation

Apart from canonical EDUs that coincide with clauses, there are also some noticeable classes of other EDUs that are not. Among these one of the common groups (11.8% of all EDUs in the Russian corpus explored in Kibrik and Podlesskaya eds. 2009) are retrospective subclausal EDUs—mostly adjuncts or attributes that semantically belong to a clause but constitute a separate short EDU following the base clause and elaborating it semantically. Consider example (2).

(2) 22.86 12 ••••(1.00) /**My** friend stood up /**behind his \desk**,
26.00 13 ••(0.15) in his /**fu-ull** \f-four \–stripes,
28.05 14 and \said:

Syntactically, EDU #13 in this example constitutes an adjunct to the clause in EDU #12, but prosodically it is clearly a separate unit. In Kibrik and Podlesskaya 2009 eds. we called this kind of retrospective subclausal EDU a **parcellation**. It emerges for the reason that the speaker has planned a clause containing too much new information, thus violating Chafe's (1994) one new idea constraint. In such situations the speaker typically chunks a clause into two pieces, conveying the adjunct as a parcellated EDU. Syntagmatically, parcellated elaborations are quite typical EDUs. But paradigmatically they present one of the most typical kinds of outliers in the segmental discourse structure.

4. Non-discrete boundaries

As was pointed out above, transcription of spoken discourse is a matter of expert judgement. That includes segmentation of the flow of talk into EDUs. An experienced transcriber takes into account all the relevant criteria and posits EDU boundaries. If two or more transcribers, working in the same fixed framework and having a comparable level of experience, analyze the same sample of talk, the instances of divergence are few. However, some divergences occur. Moreover, divergence may happen in the mind of a single transcriber as well; in other words, s/he may have doubts on whether a boundary must be posited or not, for the reason that the criteria of EDU identification are not fully consistent with each other. This happens in line #2 of example (3).

(3)	0.59	1	When I came /back,
	1.39	2	from one of those əə(0.26) ••(0.14) \trips ¦ from down to-o əə(0.27) ••(0.10) /–Cartagena,

Line #2 as a whole is a parcellated elaboration on the base clause found in line #1. However, line #2 includes symbol ¦ that indicates the location of transriber's doubts. The fragment from one of those əə(0.26) ••(0.14) \trips can be considered separately, and in such case can be taken as a fully-fledged EDU with the primary EDU accent on \trips. Under such interpretation, the subsequent fragment from down to-o əə(0.27) ••(0.10) /–Cartagena, can be interpreted as a separate EDU, functioning as a parcellated attribute of the preceding EDU. On the other hand, the fragment from down to-o əə(0.27) ••(0.10) /–Cartagena, is vocalized as if it were an immediate continuation of the preceding fragment: there is no pause, there is no reset of intonation contour at the beginning, and the accent on /–Cartagena sounds as a good candidate for a primary accent of the whole construction. This combination of considerations leads to the interpretation shown in (3): the whole construction is treated as a single EDU, but with a tentative boundary in the middle. It is important to emphasize that is not merely an issue of representation and not just a transcriber's difficulty. Rather, by providing the transcription shown in (3) we are faithful to the equivocal vocalization employed by the speaker himself. It is not just the transcriber who has doubts about positing boundaries; it was the the speaker who hesitated on whether to end his EDU on the word \trips or elaborate it further. The EDU in line #2 is thus a syntagmatic hybrid of two potentially independent EDUs.

Consider another example.

(4) 54.79 44 ••••(2.12) I /stayed in the /US –Navy ¦ ↓**s-seventeen** \ years and ten \m<u>o</u>nths.

Line #44 shown in this example sounds as having far too many accents for a single EDU. Also, if one considers the fragment ••••(2.12) I /stayed in the /US –Navy by itself, it sounds as a self-sufficient EDU with the primary accent on the word –Navy. However, it must be taken into account that the speaker talks with emphasis (especially on ↓**s-seventeen**) and uses a kind of scanning prosody (almost word-by-word accenting), which is the reason for the multiplicity of accents. In addition, the sequence –Navy ¦ ↓**s-seventeen** is vocalized without a slightest pause and with clearly continuous intonation. So, overall, the decision is made to transcribe the whole clause as a single EDU with a shadowy boundary in the middle.

Instances of non-discrete EDU boundaries are not overwhelming in spoken discourse, but they are not too rare either (perhaps one instance out of 10 or 15 EDUs). So the issue of non-discrete boundaries in speech are hardly negligible. We find that non-discrete effects, already familiar from other levels of language, are also characteristic of the discourse level, both syntagmatically and paradigmatically.

5. Conclusion

The discovery of discourse segmentation by a number of independent researchers several decades ago demonstrated that the level of discourse has units, just as any other level. However, analytic difficulties associated with EDU identification may lead some to believe that discourse is not as segmented as other levels. That would be a misguided conclusion. Non-discrete effects occur at various levels of language, including phonetics, morphology, and syntax. To quote from Edward Sapir, "unfortunately, or luckily, no language is tyrannically consistent. All grammars leak." (Sapir 1921: 38). The level of discourse structure is not exempt from non-discreteness either: we have seen examples of paradigmatic and syntagmatic deviations from the focal, or prototypical, EDUs. The existence of non-discrete effects in discourse segmentation does not undermine the very idea of segmentation, just as coarticulation does not imply that phonemes do not exist. Rather, phonemes, EDUs, and other units are not as crisp and clean as our digital mind would want them to be. In order to address linguistic reality in its actual complexity, we have to recognize that segmentation follows the principles of focal structure, which is the general property of language and cognition.

References

1. *Alexandrov Ju. I., Sergienko E. A.* (2003), Psixologicheskoe i fiziologicheskoe: kontinual'nost' ili diskretnost'? [Psychological and physiological: continuity or discreteness?] Psixologicheskij zhurnal, Vol. 24.6, pp. 98–109.

2. *Chafe W.* (1994), Discourse, consciousness, and time, University of Chicago Press, Chicago.
3. *Chance S.* (2014), The cortical microstructural basis of lateralized cognition: A review, Frontiers in Psychology, 5:820, doi: 10.3389/fpsyg.2014.00820.
4. *Cienki A.* (2015), Ponjatie dinamicheskogo diapazona kommunikativny dejstvij v teorii kognitivnoj lingvistiki [The notion of the dynamic scope of relevant behaviors in cognitive linguistic theory], in A. A. Kibrik, A. D. Koshelev, A. V. Kravchenko, Ju. V. Mazurova, and O. V. Fedorova (eds.) Jazyk i mysl': sovremennaja kognitivnaja lingvistika, Jazyki slavjanskoj kultury, Moscow, pp. 560–573.
5. *Dahl Ö.* (2001), The origin of the Scandinavian languages, in Ö. Dahl and M. Koptjevskaja-Tamm eds., The Circum-Baltic languages: Typology and contact, Vol. 1: Past and present, John Benjamins, Amsterdam, pp. 215–235.
6. *Dawkins R.* (2011), The tyranny of the discontinuous mind—Christmas 2011, https://richarddawkins.net/2013/01/the-tyranny-of-the-discontinuous-mind-christmas-2011.
7. *Engwall O.* (2000), Dynamical aspects of coarticulation in Swedish fricatives—a combined EMA & EPG study, Speech, Music and Hearing Quarterly Progress and Status Report (TMH-QPSR), Vol. 41.4, pp. 49–73.
8. *Goddard C.* (2011), Semantic analysis: A practical introduction, Oxford University Press, Oxford.
9. *Hock H. H., Joseph B. D.* (1996), Language history, language change, and language relationship: An introduction to historical and comparative linguistics, de Gruyter, Berlin.
10. *Iwasaki Sh., Tao H.* (1993), A comparative study of the structure of the intonation unit in English, Japanese, and Mandarin Chinese, paper read at the annual meeting of the Linguistics Society of America, Los Angeles, CA, Jan. 9. 1993.
11. *Janda L. A.* (2015), Aspektual'nye tipy russkogo glagola: peresmatrivaja tipologiju Krofta [Russian aspectual types: Croft's typology revised], in A. A. Kibrik, A. D. Koshelev, A. V. Kravchenko, Ju. V. Mazurova, and O. V. Fedorova (eds.) Jazyk i mysl': sovremennaja kognitivnaja lingvistika, Jazyki slavjanskoj kultury, Moscow, pp. 213–237.
12. *Kibrik A. A.* (2012a), Non-discrete effects in language, or the Critique of Pure Reason 2, Ju. I. Alexandrov et al. eds., Fifth International Conference on Cognitive Science, June 18–24, 2012, Kaliningrad, Russia, Vol. 1, pp. 81–83.
13. *Kibrik A. A.* (2012b), Prosody and local discourse structure in a polysynthetic language, Ju. I. Alexandrov et al. eds., Fifth International Conference on Cognitive Science, June 18–24, 2012, Kaliningrad, Russia, Vol. 1, pp. 80–81.
14. *Kibrik A. A.* (2013), Nediskretnost' v jazyke i fokal'naja struktura [Non-discreteness in language and focal structure], S. I. Masalova (ed.) First International Forum on Cognitive Modelling, September 14–21, 2013, Italy, Milano Marittima, Part 1, pp. 113–116.
15. *Kibrik A. A., Podlesskaya V. I. eds.* (2009), Rasskazy o snovidenijax: Korpusnoe issledovanie ustnogo russkogo diskursa [Night Dream Stories: A corpus study of spoken Russian discourse], Jazyki slavjanskix kul'tur, Moscow.

16. *Labov W.* (1973), The boundaries of words and their meanings, in C.-J. Bailey and R. Shuy (eds.), New ways of analyzing variation in English. Georgetown U. Press, Washington, DC, pp. 340–373.
17. *Lakoff G.* (1987), Women, fire, and dangerous things: What categories reveal about the mind, University of Chicago Press, Chicago.
18. *Matsumoto K.* (2003), Intonation units in Japanese conversation: Syntactic, informational, and functional structures, John Benjamins, Amsterdam.
19. *Rosch E.* (1973), Natural categories, Cognitive Psychology, vol. 4, pp. 328–350.
20. *Sapir E.* (1921), Language: An introduction to the study of speech, Harvest Books, New York.
21. *Trudgill P.* (2011), Sociolinguistic typology: Social determinants of linguistic complexity, Oxford University Press, Oxford.
22. *van Deemter K.* (2010), Not exactly: In defense of vagueness, Oxford University Press, Oxford.
23. *Wittgenstein L.* (1953/2001), Philosophical Investigations, Blackwell Publishing.
24. *Wouk F.* (2008), The syntax of intonation units in Sasak, Studies in Language, vol. 32.1, pp. 137–162.
25. *Zaliznjak A. A.* (2006), Mnogoznachnost' v jazyke i sposoby ee predstavlenija [Polysemy in language and ways to represent it], Jazyki slavjanskix kul'tur, Moscow.

РАЗРАБОТКА ФАКТОРНЫХ МОДЕЛЕЙ ЯЗЫКА ДЛЯ АВТОМАТИЧЕСКОГО РАСПОЗНАВАНИЯ РУССКОЙ РЕЧИ

Кипяткова И. С. (kipyatkova@iias.spb.su)[1,2]
Карпов А. А. (karpov@iias.spb.su)[1,3]

[1]Санкт-Петербургский институт информатики и автоматизации Российской академии наук (СПИИРАН), Санкт-Петербург, Россия

[2]Санкт-Петербургский государственный университет аэрокосмического приборостроения (ГУАП), Санкт-Петербург, Россия

[3]Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики (ИТМО), Санкт-Петербург, Россия

В статье описывается процесс создания и исследования факторных моделей языка для системы автоматического распознавания русской речи. Различные факторные модели языка и базовая 3-граммная модель были обучены на текстовом корпусе, сформированном из интернет-сайтов ряда электронных газет, содержащем более 350 млн словоупотреблений. Были созданы факторные модели с фиксированными и с параллельными путями возврата, при этом использовалось 5 лингвистических факторов: словоформа, лемма, основа слова, часть речи и метка морфологических признаков. Оптимизация параметров моделей производилась с использованием генетического алгоритма. Созданные модели были внедрены в систему автоматического распознавания русской речи и используются на этапе переоценки списка лучших гипотез распознавания. В ходе экспериментов по распознаванию слитной русской речи со сверхбольшим словарем относительное уменьшение процента неправильно распознанных слов, полученное после выполнения переоценки списка гипотез распознавания с использованием факторных моделей языка, интерполированных с базовой 3-граммной моделью, составило 8%.

Ключевые слова: факторные модели языка, автоматическое распознавание речи, русская речь, корпусные исследования

DEVELOPMENT OF FACTORED LANGUAGE MODELS FOR AUTOMATIC RUSSIAN SPEECH RECOGNITION

Kipyatkova I. S. (kipyatkova@iias.spb.su)[1,2]
Karpov A. A. (karpov@iias.spb.su)[1,3]

[1]St. Petersburg Institute for Informatics and Automation of RAS (SPIIRAS), St. Petersburg, Russia

[2]St. Petersburg State University of Aerospace Instrumentation (SUAI), St. Petersburg, Russia

[3]ITMO University, St. Petersburg, Russia

In this paper, we present a study of factored language models (FLM) of Russian for rescoring N-best lists in automatic speech recognition (ASR) systems. We used 3-gram language models as baseline. Both 3-gram and factored language models were trained on a text corpus collected from recent Internet online newspapers; total size of the text corpus is about 350 million words (2.4 Gb data). For FLM creation, we used five linguistic factors: word-form, word lemma, stem, part-of-speech, and morphological tag. We studied several FLMs with two factors (word-form plus one of the other factors) using 2 fixed backoff paths: (1) the first drop was of the most distant word and factor, then—of the less distant ones; (2) the first drop was of the words in time-distance order, then drop of the factors in the same order. We investigated the influence of a factor set and backoff paths on language model perplexity and word error rate (WER). Also we created FLMs with some parallel generalized backoff paths. Optimization of the FLM parameters was carried out by means of the genetic algorithm. The FLMs were embedded in the automatic Russian speech recognition system with a very large vocabulary. Experimental results on continuous Russian speech recognition task showed a relative WER reduction of 8% when the FLM was interpolated with the baseline 3-gram model.

Key words: factored language models, automatic speech recognition, Russian speech, corpus studies

1. Introduction

The most widely used language models (LMs) are statistical n-gram models, which estimate the probability of appearance of a word sequence $X = (W_1, W_2, ..., W_m)$ in a text [16]. Rich morphology of the Russian language leads to increasing the perplexity of n-gram models. These models are efficient for many languages, but for Russian they do not work so well. Russian is a morphologically rich inflective language. This results in the increasing of vocabulary size as well the perplexity of n-gram

language models. In [25], it was shown that changing the vocabulary size from 100K to 400K words increases the English model perplexity by 5.8% relatively, while the Russian model perplexity increases by as much as 39.5%.

A state-of-the-art alternative to *n*-gram language models is a factored language model (FLM) that for the first time was introduced in order to deal with the morphologically rich Arabic language [4]. Then it has been used for many other morphologically rich languages. This model incorporates various morphological features (factors) and it can be applied to inflective languages too. So, a word is represented as a vector of *k* factors: $w_i = (f_i^1, f_i^2, \dots, f_i^k)$. Factors of a given word can be such as word-form, morphological class, stem, root, and other grammatical features. Probabilistic language model is constructed with sets of the factors.

In [23], a FLM was incorporated at different stages of speech recognition: N-best list rescoring and recognition stage. Recognition results showed an improvement of word error rate (WER) by 0.8–1.3% with the FLM used for N-best rescoring task depending on the test speech corpus; and the usage of FLM at speech recognition gave additional improving of WER by 0.5%.

A FLM was applied for lattice rescoring in [20]. The decoder generated a lattice of 100 best alternatives for each test sentence using a word-based bigram LM with 5K vocabulary. Then the lattice was rescored with various morpheme-based and factored language models. Word recognition accuracy obtained with the baseline model was 91.60%, and the usage of the FLM increased word recognition accuracy up to 92.92%.

In [3], a morpheme-based trigram LM for Estonian was used for N-best list generating. The vocabulary of the language model consisted of 60K word particles. Recognized morpheme sequences were reconstructed to word sequences. A FLM, which used words and their part-of-speech (POS) tags, was applied to rescore N-best hypotheses. A relative WER improvement of 7.3% was obtained on a large vocabulary.

FLMs are also used for code-switching speech [1, 9]. In [1], for code-switching speech the following factors were analyzed: words, POS tags, open class words, and open class word clusters. FLM was used at the speech decoding stage. For this purpose BioKIT speech decoder [21] was extended to support such models. Experiments on recognition of Mandarin-English code-switching speech showed a relative reduction of mixed error rate by 3.4%. In [2], a FLM was combined with recurrent neural networks (RNN) for Mandarin-English code-switching language modeling task. The combined LM gave a relative improvement of 32.7% comparing to the baseline 3-gram model.

An application of FLMs for Russian speech recognition is described in [22]. The FLM was trained on a text corpus containing 10M words with a vocabulary size of about 100K words. FLMs were created using the following factors: word, lemma, morphological tag, POS, and gender-number-person factor. TreeTagger tool [17] was used for obtaining linguistic factors. Influence of different factors and backoff paths on the perplexity and WER was tested. FLM was used for rescoring 500-best lists. Evaluation experiments showed that FLM allows achieving 4.0% WER relative reduction, and 6.9% relative reduction was obtained after interpolation of the FLM with the baseline 3-gram model.

2. Creation of Factored Language Models for Russian

There are two main issues at development of FLM [14]:

1. Choosing an appropriate set of factor definitions using data-driven techniques or linguistic knowledge.
2. Finding the best statistical model for these factors.

One of the problems at creating statistical LMs is the lack of training data (especially for under-resourced languages) [9]. To solve this problem backoff methods are used [16]. In word *n*-gram modeling, backing-off is performed by dropping first the most distant word, followed by the second most distant word, and so on until the unigram language model is used. This process is illustrated in Figure 2(a). In FLM, there is no obvious path of backoff [4]. In FLMs, any factor can be dropped at each step of the backoff process, and it is not obvious, which factor to drop first. In this case, several backoff paths are possible, that results in a backoff graph. An example of the backoff graph is presented in Figure 1(b). The graph shows all possible single step backoff paths, where exactly one variable is dropped per each step.

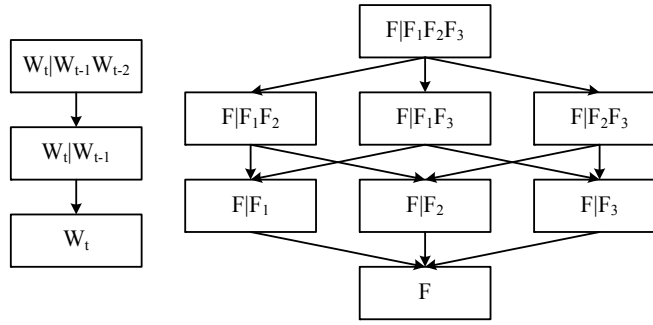


Fig. 1. Backoff graphs for ***n***-gram and FLMs: (a) backoff path for a 3-gram language model over words; (b) backoff graph with three parent factors F1, F2, F3

In order to choose the best factor set and backoff path, linguistic knowledge or data-driven techniques can be applied. In [23], it was shown that an automatic method that uses Genetic Algorithm (GA) for optimization of the factor set, backoff path, and smoothing techniques, performs better than the manual search in terms of perplexity. The goal of this method is to find a combination of parameters that produces a FLM with a low perplexity on unseen test data [14].

For the language model creation, we collected and automatically processed a Russian text corpus of some on-line newspapers. It contains news texts of different topics: politics, economy, culture, sport, etc. The procedure of preliminary text processing and normalization is described in [8]. The size of the corpus after text normalization and deletion of doubling and short (<5 words) sentences is over 350M words, as well as it contains above 1M unique word-forms.

The software “VisualSynan” of AOT project [18] was used for obtaining morphological features for words. This tool can make a morphological analysis of Russian, English, and German texts. Output of the morphological analysis is quite correct, although some errors are exist. We used 5 linguistic factors: word-form, its lemma, stem, part-of-speech (POS), and morphological tag. The training text corpus was processed to replace words with their factors. For example, the word-form ‘**схеме**’ (“scheme”) is replaced with the vector {W-**схеме**: L-**схема**: S-**схем**: P-**сущ**: M-bc}, where W means a word-form, L denotes a lemma, S is a stem, P is POS, M is a morphological tag, which indicates that the given word-form is a noun with feminine gender, singular, dative case.

2.1. FLMs with fixed backoff paths

We used the SRI Language Modeling Toolkit (SRILM) [19] for LM creation. At first, we created 2-factor LMs with the word-form plus one of the other factors. To create these models we used two fixed backoff paths:

1. The first drop was of the most distant word-form and factor, then—of the less distant ones (Figure 1a).
2. The first drop was of the word-forms in time-distance order, and then the drop of the factors in the same order (Figure 2a).

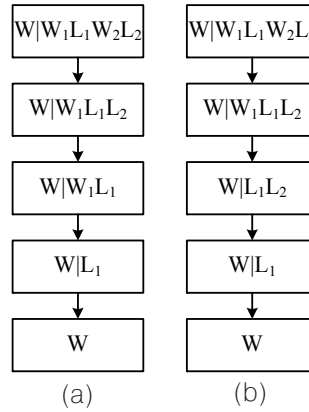


Fig. 2. Backoff paths for WL model: (a) backoff path 1; (b) backoff path 2

For example, for the real trigram “**вагонов грузового состава**” (“wagons of freight train”) the backoff path is the following:

вагонов грузового состава
грузового состава
состава

When creating 2-factor LM (word-form and lemma) this trigram is converted into a sequences of factors: “W-**вагонов** L-**вагон** W-**грузового** L-**грузовой** W-**состава** L-**состав**”. Backoff paths can be the following:

Backoff path 1:

L-вагон W-вагонов L-грузовой W-грузового W-состава
L-вагон L-грузовой W-грузового W-состава
L-грузовой W-грузового W-состава
L-грузовой W-состава
W-состава

Backoff path 2:

L-вагон L-грузовой W-вагонов W-грузового W-состава
L-вагон L-грузовой W-грузового W-состава
L-вагон L-грузовой W-состава
L-грузовой W-состава
W-состава

When creating LMs, discounting techniques are used to assign nonzero probabilities to *n*-grams that were not observed in the training corpus by discounting probabilities of the observed *n*-grams [24]. Therefore, we investigated FLMs with different discounting techniques: (1) Good-Turing; (2) Unmodified Kneser-Ney; (3) Modified Kneser-Ney; (4) Witten-Bell; (5) Natural [24]. Perplexities of the created FLMs are shown in Table 1; they were calculated on text data consisting of phrases (33M word usage in total) from another online newspaper "Фонтанка.ru" (www.fontanka.ru), which was not used for LM training. The models built with backoff path 1 have smaller perplexities for all discounting techniques and factors. Some discounting techniques gave better results depending on factor combinations. The best perplexity was obtained using the LM with word-form and lemma factors created with the modified Kneser-Ney discounting technique. Also this discounting method gave better (smaller) perplexity for all other LMs excepting the model with word-form and part-of-speech factors. For this model the Good-Turing discounting method was the best. The largest (worst) LM perplexity was obtained using the model with word-form and stem factors with the Witten-bell discounting. The perplexity of the baseline 3-gram LM was 553 [10].

Table 1. Perplexity of FLMs with different discounting techniques and backoff paths

Factors	Discounting techniques									
	Good-Turing		Unmodified Kneser-Ney		Modified Kneser-Ney		Witten-Bell		Natural	
	Path 1	Path 2	Path 1	Path 2	Path 1	Path 2	Path 1	Path 2	Path 1	Path 2
WM	573	696	593	724	566	691	749	898	761	916
WL	557	597	550	603	**529**	577	826	1007	747	779
WP	572	636	649	755	623	729	725	727	734	762
WS	617	685	617	701	595	672	879	1098	824	895

2.2. FLMs with parallel generalized backoff

We created also FLM using all factors and parallel generalized backoff. Since the creation of such a model requires a large amount of memory, we used only a part of the text corpus, which contains 100M words. We applied the genetic algorithm (GA) [5] to find the best backoff graph. As initial factors we used all mentioned above factors and discounting methods, as well as the time context of 2. GA was implemented using the population size of 10 and the maximum number of generation of 20 [11].

We chose two models, which are the best in the terms of perplexity for the experiments on Russian ASR. A backoff graph for the first model (FLM 1) is presented in Figure 3; the backoff graph for the second model (FLM 2) is presented in Figure 4. In these figures, a digit after a factor symbol denotes a time context.

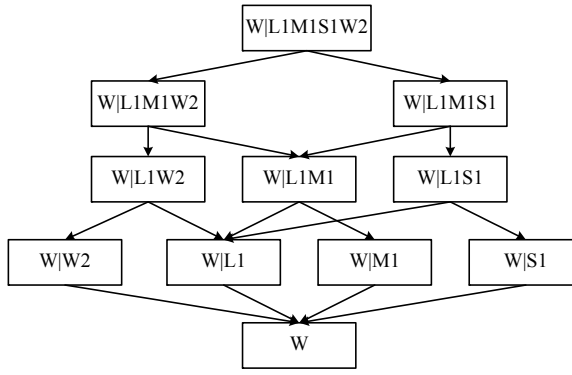


Fig. 3. Backoff graph for FLM 1

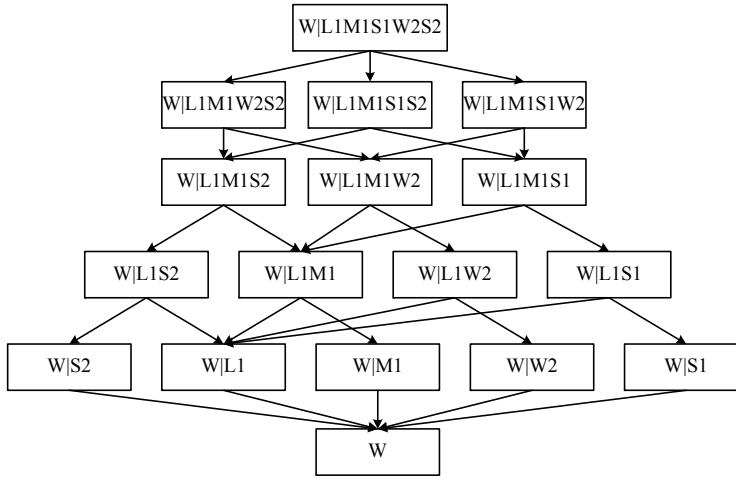


Fig. 4. Backoff graph for FLM 2

Both models use 4 factors: lemma, morphological tag, stem, word-form, and three discounting methods on different stages of backing-off: Unmodified Kneser-Ney, Modified Kneser-Ney, and Witten-Bell. The perplexity of FLM 1 is 589, and the perplexity of FLM 2 is 618.

3. Russian Speech Recognition System with FLM

Architecture of the Russian ASR system with developed FLMs is presented in Fig. 5.

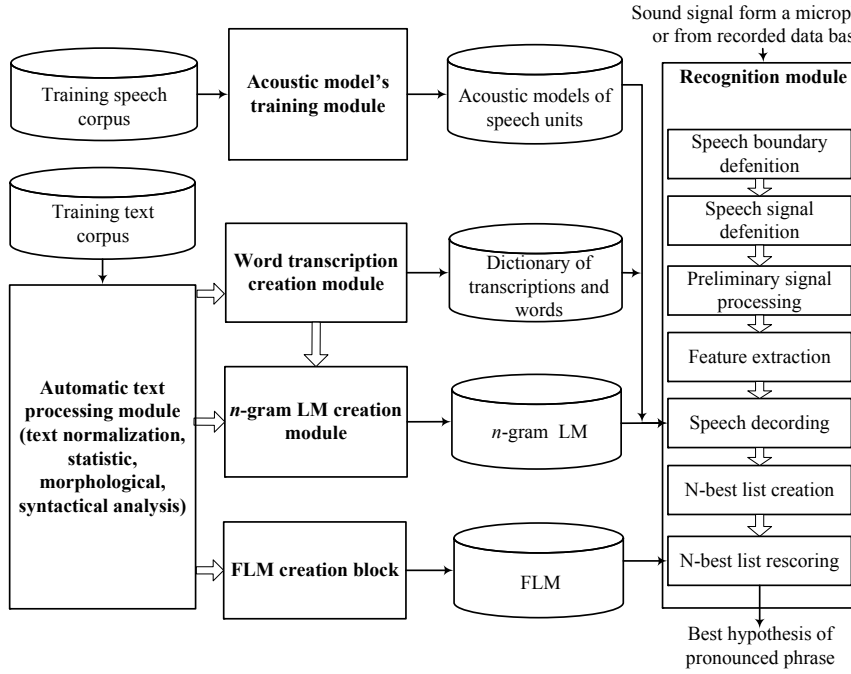


Fig. 5. Architecture of Russian ASR system with FLM

The system works in 2 modes [12]: training and recognition. In the training mode, acoustic models of speech units, a phonemic vocabulary of word-forms, as well as *n*-gram and factored LMs are created. In the speech recognition mode, an input speech signal is transformed into the sequence of feature vectors (Mel-Frequency Cepstral Coefficients with the 1st and 2nd order derivatives are used), and then the search of most probable hypotheses is performed with the help of preliminary trained acoustic and language models. FLM is used at the stage of post-processing for N-best list rescoring. Thereby, on the speech recognition stage, 3-gram LM is used for creating N-best list and then FLM is applied for rescoring obtained N-best list of hypotheses and for selection of the best recognition hypothesis for pronounced phrase.

4. Experiments on continuous Russian speech recognition

4.1. Training and testing speech corpora

For training the speech recognition system we used our own corpus of spoken Russian speech, created by SPIIRAS in 2008–2009 in the framework of Euronounce project [6, 7]. The speech data were collected in clean acoustic conditions, with 44.1 kHz sampling rate, 16-bit audio quality. The signal-to-noise ratio (SNR) of 35–40 dB at least was provided. The database consists of 16,350 utterances pronounced by 50 native Russian speakers (25 male and 25 female). Each speaker pronounced more than 300 phonetically-balanced and meaningful phrases. Total duration of the speech data is about 21 hours.

Acoustic models were created with the help of HTK toolkit [26]. As for acoustic features, we used 13-dimentional MFCCs with the 1st and 2nd order derivatives calculated from the 26-channel filter bank analysis of 20 ms long frames with 10 ms overlap. Cepstral mean subtraction (CMS) is applied to audio feature vectors. For acoustic modeling, continuous density Hidden Markov Models (HMM) were used, and each phoneme was modeled by one HMM.

To test the system we used a speech corpus that contains 500 phrases pronounced by 5 speakers (each speaker pronounced the same 100 phrases). The phrases were taken from the materials of the on-line newspaper "Фонтанка.ru" that was not used in the training data.

4.2. Study of FLMs with fixed backoff paths

Russian ASR system was built on the base of Julius ver. 4.2 decoder [15]. System's performance was estimated by the word error rate (WER) measure. At the speech decoding stage, 3-gram LM was used. WER obtained with this model was 26.54% [10]. The vocabulary size was 150K words. The out-of-vocabulary rate for the test set was 1.1%. The baseline ASR system produced 20-best lists of hypotheses for each pronounced phrase. The rescoring of the 20-best lists was carried out using created FLMs. The recognition results are summarized in Table 2.

Table 2. WER obtained after 20-best list rescoring (%)

FLMs	Discounting techniques									
	Good-Turing		Unmodified Kneser-Ney		Modified Kneser-Ney		Witten-Bell		Natural	
	Path 1	Path 2	Path 1	Path 2	Path 1	Path 2	Path 1	Path 2	Path 1	Path 2
WM	27.87	28.00	27.79	28.09	28.15	28.16	**27.30**	27.55	27.58	27.40
WL	28.45	28.78	28.37	28.82	28.28	28.99	27.83	28.39	27.88	28.67
WP	28.61	28.61	28.58	28.71	28.63	28.88	27.72	28.48	28.33	28.91
WS	29.93	30.24	29.78	30.19	30.02	30.28	29.01	29.46	28.90	29.91

FLMs	Discounting techniques									
	Good-Turing		Unmodified Kneser-Ney		Modified Kneser-Ney		Witten-Bell		Natural	
	Path 1	Path 2	Path 1	Path 2	Path 1	Path 2	Path 1	Path 2	Path 1	Path 2
Interpolated models										
WM+3-gram	25.00	24.89	24.57	24.93	**24.44**	24.78	24.94	25.22	25.41	25.36
WL+3-gram	25.51	25.54	25.54	25.67	25.58	25.43	25.51	25.47	24.98	25.47
WP+3-gram	25.21	25.28	25.30	25.32	25.07	25.24	25.47	25.64	25.60	25.43
WS+3-gram	25.97	25.92	26.03	25.86	25.88	25.90	26.05	25.86	25.49	25.90

We produced lists of 20-best hypotheses and rescored them using created FLMs. The best results were obtained using the LM with word-form and morphological tag factors created with the Witten-Bell discounting. Optimal WER value was 27.30%. So, the WER was worse than one obtained before N-best list rescoring. For models with other combination of factors the Witten-Bell discounting also gave better results, although in terms of perplexity this discounting method was not the best. Then we carried out linear interpolation of FLMs with the baseline 3-gram LM. The lowest WER=24.44% was obtained after interpolation of the baseline model with the FLM, in which word-form and morphological factors were used. This model was created using modified Kneser-Ney discounting technique with the backoff path 1.

Then, we produced N-best lists with the number of hypotheses from 10 to 50 and performed their rescoring using FLM with the Modified Kneser-Ney discounting technique interpolated with 3-gram model. Recognition results are presented in Table 3. Also in the table oracle WER, which is minimal value of WER that can be obtained choosing the most accurate hypothesis from N-best list, is shown. From the table we can see that rescoring of 20-best list gives better results.

Table 3. WER obtained after rescoring of N-best lists (%)

Language models	N=10		N=20		N=50	
	Path 1	Path 2	Path 1	Path 2	Path 1	Path 2
3-gram (oracle WER)	18.52		16.63		15.34	
3-gram + WM	24.83	24.94	**24.44**	24.78	24.55	24.66
3-gram + WL	25.79	25.71	25.58	25.43	25.60	25.37
3-gram + WP	25.43	25.54	25.07	25.24	25.15	25.26
3-gram + WS	25.82	26.01	25.88	25.90	25.90	26.10

4.3. Study of LMs with parallel generalized backoff

Then experiments on rescoring 20-best lists using FLMs with the parallel generalized backoff method were conducted. Obtained results are presented in Table 4. The use

of FLMs for N-best list rescoring did not improve the ASR results. Therefore, we have performed a linear interpolation of FLMs with the baseline model. The best WER was obtained with the FLM 1 interpolated with the baseline 3-gram LM (WER=24.53%).

Table 4. WER obtained after rescoring 20-best lists with parallel generalized backoff (%)

Language models	WER, %
3-gram	26.54
FLM 1	27.94
FLM 2	28.56
FLM 1 + 3-gram	**24.53**
FLM 2 + 3-gram	24.74

Figure 5 shows the 20-best list of ASR for the Russian phrase: "Основой нашего эфира станет мировая музыкальная классика во всем многообразии жанров, стилей и направлений» ("The base of our broadcast will become world classical music in variety of genres, styles, and trends"). The hypotheses are ranked according to descending probability. After rescoring of this 20-best list using FLM 1 interpolated with the baseline 3-gram LM, the hypothesis #4 was selected as the best one. So, after N-best list rescoring we obtained the correct hypothesis for this utterance.

#1 <s> мы заранее договорились что разговор нужны для публикации проста надо познакомиться поближе </s>
#2 <s> мы заранее договорились что разговор наш не для публикации проста надо познакомиться поближе </s>
#3 <s> мы заранее договорились что разговор нужны для публикаций портала познакомиться поближе </s>
#4 <s> мы заранее договорились что разговор наш не для публикации просто надо познакомиться поближе </s>
#5 <s> мы заранее договорились что разговор наш не для публикации портала познакомиться поближе </s>
#6 <s> мы заранее договорились что разговор нужны для публикаций просто надо познакомиться поближе </s>
#7 <s> мы заранее договорились что разговор нужны для публикаций проста надо познакомиться поближе </s>
#8 <s> мы заранее договорились что разговор наш мир для публикации проста надо познакомиться поближе </s>
#9 <s> мы заранее договорились что разговор нож не для публикации просто надо познакомиться поближе </s>
#10 <s> мы заранее договорились что разговор нашли для публикации проста надо познакомиться поближе </s>
#11 <s> мы заранее договорились что разговор нож не для публикации портала познакомиться поближе </s>
#12 <s> мы заранее договорились что разговор наш ни для публикации проста надо познакомиться поближе </s>
#13 <s> мы заранее договорились что разговор наш мир для публикации просто надо познакомиться поближе </s>
#14 <s> мы заранее договорились что разговор нужные для публикации проста надо познакомиться поближе </s>
#15 <s> мы заранее договорились что разговор наш не для публикаций портала познакомиться поближе </s>
#16 <s> мы заранее договорились что разговор наш мир до публикации проста надо познакомиться поближе </s>
#17 <s> мы заранее договорились что разговор нужный для публикации проста надо познакомиться поближе </s>
#18 <s> мы заранее договорились что разговор на шнидер публикации проста надо познакомиться поближе </s>
#19 <s> мы заранее договорились что разговор нашли для публикации просто надо познакомиться поближе </s>
#20 <s> мы заранее договорились что разговор наш ни для публикаций портала познакомиться поближе </s>

Fig. 6. An example of N-best list of recognition hypotheses

Table 4 shows that the WER obtained after applying the LMs with parallel backoff paths slightly increased comparing to results obtained after applying the models with fixed backoff paths. The reason for this is that models with parallel backoff paths were trained on a portion of the corpus (100M word usage). The disadvantage of FLMs with many factors and parallel backoff paths is that these models require a large amount of memory; in our case training these models required 64 Gb RAM memory. However,

it is possible to obtain decreasing WER even by training these models using a small train corpus that is an obvious advantage of FLMs.

Our experimental results are consistent with those obtained in [22], but we used another morphological parser—AOT [18] instead of TreeTagger [17]. For our experiments we used the training text corpus of 350 million words that is in 35 times larger than the set in [22]. Moreover, our WER results are better than reported in [22], and they confirm the hypothesis that the use of FLM for N-best list rescoring improves recognition accuracy. Also we can conclude that we obtained a larger relative reduction of WER in comparison with some other researches for other languages (for example, reported in [3, 20, 23]).

5. Conclusion

The study of FLMs showed that the inclusion of addition linguistic information in language models can improve the performance of ASR systems. In this paper, we compared different factor sets in terms of the word error rate. We obtained relative WER reduction of 8% comparing to the baseline ASR system. In further research, we plan to investigate FLMs with other factors as well as other types of statistical language models.

This research is partially supported by the Council for Grants of the President of Russia (Projects No. MK-5209.2015.8 and MD-3035.2015.8), by the Russian Foundation for Basic Research (Projects No. 15-07-04415 and 15-07-04322), and by the Government of the Russian Federation (Grant No. 074-U01).

References

1. *Adel H., Kirchhof K., Telaar D., Vu N. T., Schlippe T., Schultz T.* (2014), Features for factored language models for code-switching speech, Proceedings of 4th International Workshop on Spoken Language Technologies for Under-resourced languages (SLTU-2014), St. Petersburg, Russia, pp. 32–38.
2. *Adel H., Vu N. T., Schultz T.* (2013), Combination of Recurrent Neural Networks and Factored Language Models for Code-Switching Language Modeling, Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL 2013), Sofia, Bulgaria.
3. *Alumae T.* (2006), Sentence-adapted factored language model for transcribing Estonian speech, Proceedings of International Conference on Acoustics, Speech and Signal Processing (ICASSP- 2006). Toulouse, France, pp. 429–432.
4. *Bilmes J. A., Kirchhoff K.* (2003), Factored language models and generalized parallel backoff, Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology, Stroudsburg, USA, Vol. 2, pp. 4–6.
5. *Gladkov L. A., Kureychik V. V., Kureychik V. M.* (2006), Genetic algorithms [Geneticheskie algoritmy]. 2nd ed., Fizmatlit, Moscow.

6. *Jokisch O., Wagner A., Sabo R., Jaeckel R., Cylwik N., Rusko M., Ronzhin A., Hoffmann R.* (2009), Multilingual speech data collection for the assessment of pronunciation and prosody in a language learning system, Proceedings of SPECOM'2009, St. Petersburg, Russia, pp. 515–520.
7. *Karpov A., Kipyatkova I., Ronzhin A.* (2011), Very Large Vocabulary ASR for Spoken Russian with Syntactic and Morphemic Analysis, Proceedings of International Conference INTERSPEECH-2011, Florence, Italy, pp. 3161–3164.
8. *Karpov A., Markov K., Kipyatkova I., Vazhenina D., Ronzhin A.* (2014), Large vocabulary Russian speech recognition using syntactico-statistical language modeling, Speech Communication, Vol. 56. pp, 213–228.
9. *Karpov A., Verkhodanova V.* (2015), Speech Technologies for Under-Resourced Languages of the World [Rechevye tehnologii dlja maloresursnyh jazykov mira] // Problems of Linguistics [Voprosy Jazykoznanija], Vol. 2, pp. 117–135.
10. *Kipyatkova I., Karpov A.* (2013), Lexicon Size and Language Model Order Optimization for Russian LVCSR // Springer International Publishing Switzerland. M. Zelezny et al. (Eds.): SPECOM 2013, LNAI 8113, pp. 219–226.
11. *Kipyatkova I., Karpov A.* (2014), Study of Morphological Factors of Factored Language Models for Russian ASR // Springer International Publishing Switzerland. A. Ronzhin et al. (Eds.): SPECOM 2014, LNAI 8773, pp. 451–458.
12. *Kipyatkova I., Karpov A., Verkhodanova V., Zelezny M.* (2014), Modeling of Pronunciation, Language and Nonverbal Units at Conversational Russian Speech Recognition // International Journal of Computer Science and Applications, Vol. 10, N 1, pp. 11–30.
13. *Kipyatkova I., Verkhodanova V., Karpov A.* (2014), Rescoring N-best lists for Russian speech recognition using factored language models, Proceedings of 4th International Workshop on Spoken Language Technologies for Under-resourced languages (SLTU-2014), St. Petersburg, Russia, pp. 81–86.
14. *Kirchhoff K., Bilmes J., Duh K.* (2007), Factored Language Models Tutorial, Tech. Report UWEETR-2007-0003, Department of Electrical Engineering, University of Washington.
15. *Lee A., Kawahara T.* (2009), Recent Development of Open-Source Speech Recognition Engine Julius, Proceedings of Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC 2009), Sapporo, Japan, pp. 131–137.
16. *Moore G. L.* (2001), Adaptive Statistical Class-based Language Modelling, PhD thesis, Cambridge University.
17. *Schmid H.* (1994), Probabilistic part-of-speech tagging using decision trees, Proc. International Conference on New Methods of Language Processing, Manchester, UK, pp. 44–49.
18. *Sokirko A. V.* (2004), Morphological modules on www.aot.ru [Morfologicheskie moduli na sajte www.aot.ru], Proceedings of International Conference "Dialogue-2004" [Trudy Mezhdunarodnoj konferencii "Dialog-2004"], pp. 559–564.
19. *Stolcke A., Zheng J., Wang W., Abrash V.* (2011), SRILM at Sixteen: Update and Outlook, Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop ASRU'2011. Waikoloa, Hawaii, USA.

20. *Tachbelie M. Y., Teferra Abate S., Menzel W.* (2009), Morpheme-based language modeling for Amharic speech recognition, Proceedings of the 4th Language and Technology Conference, LTC-2009, Posnan, Poland, pp. 114–118.
21. *Telaar D., Wand M., Gehrig D., Putze F., Amma C., Heger D., Vu N. T., Erhardt M., Schlippe T., Janke M., Herff C., Schultz T.* (2014), BioKIT — Real-time decoder for biosignal processing, Proceedings of Interspeech-2014, Singapore, pp. 2650–2654.
22. *Vazhenina D., Markov K.* (2013), Factored Language Modeling for Russian LVCSR, Proceedings of International Joint Conference on Awareness Science and Technology & Ubi-Media Computing, Aizu-Wakamatsu city, Japan, pp. 205–210.
23. *Vergyri D., Kirchhoff K., Duh K., Stolcke A.* (2004), Morphology-Based Language Modeling for Arabic Speech Recognition, Proceedings of ICSLP 2004, pp. 2245–2248.
24. *Whittaker E. W. D.* (2000), Statistical Language Modelling for Automatic Speech Recognition of Russian and English, PhD thesis, Cambridge University.
25. *Whittaker E. W. D., Woodland P. C.* (2003), Language modelling for Russian and English using words and classes, Computer Speech and Language, Vol. 17, pp. 87–104.
26. *Young S.* et al. (2009), The HTK book, Cambridge Univ. Press.

РУССКИЙ ЛЕКСИКОГРАФИЧЕСКИЙ ЛАНДШАФТ: ИСТОРИЯ О 12 СЛОВАРЯХ

Киселёв Ю. А. (ykiselev.loky@gmail.com)[1, 2]
Крижановский А. А. (andrew.krizhanovsky@gmail.com)[3]
Браславский П. И. (pbras@yandex.ru)[1, 4]
Меньшиков И. Л. (unkmas@gmail.com)[1]
Мухин М. Ю. (mfly@sky.ru)[1]
Крижановская Н. Б. (nataly@krc.karelia.ru)[3]

[1]Уральский федеральный университет, Екатеринбург, Россия
[2]Яндекс, Екатеринбург, Россия
[3]ИПМИ КарНЦ РАН, Петрозаводск, Россия
[4]Kontur Labs, Екатеринбург, Россия

Ключевые слова: лексический ресурс, словарь, тезаурус, ворднет, русский язык

RUSSIAN LEXICOGRAPHIC LANDSCAPE: A TALE OF 12 DICTIONARIES

Yuri Kiselev (ykiselev.loky@gmail.com)[1, 2]
Andrew Krizhanovsky (andrew.krizhanovsky@gmail.com)[3]
Pavel Braslavski (pbras@yandex.ru)[1, 4]
Ilya Menshikov (unkmas@gmail.com)[1]
Mikhail Mukhin (mfly@sky.ru)[1]
Nataly Krizhanovskaya (nataly@krc.karelia.ru)[3]

[1]Ural Federal University, Ekaterinburg, Russia
[2]Yandex, Ekaterinburg, Russia
[3]Institute of Applied Mathematics Research,
Karelian Research Center of RAS, Petrozavodsk, Russia
[4]Kontur Labs, Ekaterinburg, Russia

The paper reports on quantitative analysis of 12 Russian dictionaries at three levels: 1) headwords: the size and overlap of word lists, coverage of large corpora, and presence of neologisms; 2) synonyms: overlap of synsets in different dictionaries; 3) definitions: distribution of definition lengths and numbers of senses, as well as textual similarity of same-headword definitions in different dictionaries. The total amount of data in the study is 805,900 dictionary entries, 892,900 definitions, and 84,500 synsets. The study reveals multiple connections and mutual influences between dictionaries, uncovers differences in modern electronic vs. traditional printed resources, as well as suggests directions for development of new and improvement of existing lexical semantic resources.

Key words: lexical resource, dictionary, thesaurus, wordnet, Russian language

1. Introduction

The problem of analysis and comparison of existing lexical resources for Russian has arisen within the Yet Another RussNet (YARN) project[1]. YARN aims at creating an open thesaurus for Russian using crowdsourcing while maximizing the use of existing lexical-semantic resources (LSRs) [3]. From a linguistics point of view, YARN has rather traditional structure introduced in Princeton WordNet (PWN) [11] and adopted by its numerous successors and variants. YARN consists of synsets—groups of near-synonyms corresponding to a concept; synsets are linked to each other, primarily via hierarchical hyponymic/hypernymic relationships. The project is ongoing and expected to cover Russian nouns, verbs, and adjectives. The main difference from the previous projects is that it is based on crowdsourcing. We hope that crowdsourcing approach will make it possible to create a resource of satisfactory quality and size in foreseeable future and with limited financial resources. Our optimism is based both on international practice and recent examples of successful Russian NLP projects driven by volunteers.

The input information (synonymy and hierarchical relationships) to be validated by the "crowd" is a result of automatic processing of corpus and dictionary data. A brief description of the data sources and online tool that are used in the project at the moment can be found in [4].

The goal of this study is to create an inventory of available LSRs for Russian, to figure out how they relate to each other, what "gaps" in the description of Russian lexis exist and how data at hand can be incorporated into YARN. A big advantage to the study is that a large number of initially printed dictionaries are available today in machine-readable form[2]. As far as we know, no large-scale quantitative comparison of the body of Russian dictionaries has been conducted yet. We hope that our findings will be useful not only within YARN project, but also of interest for a wide lexicographic community as well.

For the study, we employed electronic versions of six printed explanatory dictionaries and three dictionaries of synonyms, online Russian Wiktionary, as well as electronic thesauri RuThes and Russian WordNet. The total amount of data in the study is 805,900 dictionary entries; 892,900 definitions, and 84,500 synsets. Despite the impressive amount of data used in the study, it still remains incomplete: not all Russian dictionaries that we would like to include in the study are available in machine-readable format, and we were not ready to conduct the whole routine of scanning, recognition, and post-processing. Moreover, available resources vary significantly in quality—both because of the structure and print layout of dictionary entries and the quality of recognition and subsequent processing (for example, we could not perform definitions analysis in one of the sources since it was impossible to parse it correctly).

We investigated the dictionary data at three levels: 1) headwords: size and overlap of headword lists, coverage of large corpora, and presence of neologisms; 2) synonyms: we attempted to align the meanings of synsets in different sources and analyze their intersections; 3) definitions: distribution of definition lengths and number of senses, as well as textual similarity of same-headword definitions in different dictionaries.

---

[1] http://russianword.net

[2] http://nlpub.ru/Ресурсы

2. Related work

In our study, we compare headword lists from different dictionaries, corpora coverage by respective word lists, make an attempt to directly compare synonym data contained in different dictionaries, as well as analyze various properties of definitions and their inter-dictionary similarity. First studies on automated analysis of dictionary data in machine-readable format can be dated back to 1980s. For example, an early paper [22] studied word frequency and length distributions of definitions in an English dictionary, distributions of semantic and part-of-speech marks, as well as coverage of definitions by the top-frequency words. Michiels and Yoshida [25, 31] proposed methods for identification of hierarchical relations between word senses based on dictionary data. Automatic thesaurus construction using existing dictionaries became widespread when open collaborative projects, primarily Wikipedia and Wiktionary, matured and accumulated sufficient data volumes. The latest example of a large multilingual thesaurus based on open data is BabelNet [27]. The current Babelnet version claims to comprise more than 40 mln glosses in 271 languages that form more than 13 mln synsets (http://babelnet.org/stats).

The work by Meyer and Gurevych [24] is probably closest to ours. The main objective of the study was to compare collaboratively constructed language resources with traditional expert-built resources. The authors juxtaposed three different language editions of Wiktionary (English, German, and Russian) and corresponding thesauri—PWN, GermaNet, and Russian Wordnet. The paper presented basic statistics of resources—the total number of headwords, parts-of-speech and senses distributions, coverage of core vocabulary and neologisms in respective languages, overlap of headword lists, as well as presence of domain and register marks. The study did not analyze definitions and synonymy information presented in both kinds of resources.

A problem closely related to our research is sense alignment, i.e. matching of identical or similar senses in different LSRs. For example, an early work [20] compared PWN and printed dictionaries based on manual coding of meanings of 18 English verbs. Current approaches use fully automated methods: for example, Matuschek and Gurevych [23] combined graph-based distances between senses with textual similarity of definitions for aligning senses between Wiktionary and Wikipedia in English and German (the study also contains a nice overview of sense alignment methods and approaches). The paper [16] describes a task-oriented comparison (such as word and sentence relatedness problems) of synonymy information presented in PWN and different editions of Roget's thesaurus.

Large corpora are widely used for building modern dictionaries, in particular—to compile and update glossaries, extract collocations, and provide word usage examples. For example, Geyken and Lemnitzer [13] used Google Books Ngram Corpus to compile a wordlist for a new dictionary of German. A survey of corpus tools for lexicography can be found in [17]. In our study we handle an inverse problem: we investigate how existing dictionaries cover corpora, as well as how neologisms extracted from temporarily labeled subcorpora are presented in lexicographic resources.

Based on the literature review, we can conclude that our study is unprecedented in number of resources involved, volumes of data processed and aspects of dictionary data analyzed. Due to large volumes and wide diversity of data we employ mainly shallow processing techniques in our study.

3. Data

The resources in the study and their quantitative characteristics with brief descriptions are shown in Tables 1a and 1b (the editions of the printed dictionaries corresponding to the analyzed electronic version are specified).

Table 1a. Summary of lexical resources in the study: descriptions of dictionaries

Resource	Title [reference], year of the first edition	Editor(s)	Brief description and individual features
Explanatory dictionaries of classical type			
USH	Explanatory Dictionary of the Russian Language [9], 1935	D. N. Ushakov	influence of the Soviet ideology on definitions and examples; detailed system of style labels; obsolescence of the whole dictionary
OZH	Explanatory Dictionary of the Russian Language [10], 1949 (1992)	S. I. Ozhegov, N. Yu. Shvedova	popular normative dictionary; core vocabulary of the Russian literary language; brief examples
MAS	Small Academy Dictionary (Dictionary of the Russian language) [8], 1957	A. P. Evgenyeva	scientific approach, definitions with high accuracy; specific presentation of shades of meaning (à reduced number of isolated meanings); large number of usage examples
BTS	Big Dictionary of the Russian Language [14], 1998	S. A. Kuznetsov	MAS successor with a significantly extended word list; concise layout due to space limitations (one volume)
EFR	New dictionary of Russian [28], 2000	T. F. Efremova	large word list; extended number of meanings; systematic representation of regular polysemy; a large number of morphemes and MWEs; tendency to scientific definitions; no usage examples
ZLZ	Russian Grammar Dictionary [32], 1977	A. Zaliznyak	grammar dictionary (no definitions); one of the largest wordlists in the Russian lexicography by the time of first edition; the basis of almost all Russian lemmatizers
Synonym dictionaries			
ABR	Russian dictionary of synonyms and semantically similar expressions [1], 1900	N. Abramov	the oldest resource in the study, often used for Russian NLP
EVG	Dictionary of synonyms [6], 1970	A. Evgenyeva	large word list; significant number of usage examples, relies on the same initial data as BTS and MAS
BAB	Dictionary of synonyms of the Russian Language [7], 2011	L. Babenko	modern ideographic thesaurus

Resource	Title [reference], year of the first edition	Editor(s)	Brief description and individual features
Electronic lexical resources			
RWN	Russian Wordnet (http://wordnet.ru) [12], 2003		automatic translation of approx. 45% of PWN synsets based on parallel corpus, bilingual dictionaries and dictionaries of synonyms
WIKT	Machine-readable Wiktionary (http://ru.wiktionary.org) based on data from Russian Wiktionary [18], 2004		free multilingual online dictionary and thesaurus that can be collaboratively edited by users
RUT	Thesaurus RuThes-lite (http://www.labinform.ru/pub/ruthes) [21], 2014		linguistic ontology consisting of concepts and their relationships; same-root words (different POS) can belong to the same concept; concepts provided with definitions from WIKT

Table 1b. Summary of lexical resources in the study: quantitative characteristics (the values in parentheses in columns 3 and 4 correspond to synsets)[3]

Resource	# of entries, *10^3	# of unique lexical units, *10^3	# of MWE, *10^3	# of defs, *10^3
Explanatory dictionaries				
USH	88.8	87.1	0.0	130.5
OZH	41.2	40.3	0.0	n/a
MAS	83.5	81.6	0.0	135.8
BTS	76.3	103.2	0.0	111.8
EFR	135.2	123.7	2.3	219.0
ZLZ	93.4	93.4	0	0
Synonym dictionaries				
ABR	5.4	5.4 (16.0)	0.0 (2.1)	0
EVG	5.5	4.6 (16.4)	0.0 (0.3)	n/a
BAB	5.0	5.1 (19.6)	0.0 (1.2)	5.0
Electronic lexical resources				
RWN	51.7	30.8	9.3	74.6[3]
WIKT	193.5	192.0	5.8	161.2
RUT	26.4	96.7	46.6	54.9

All dictionary data were converted to a uniform machine-readable representation. For each entry we kept headword (with variations), definitions, and synonyms. Headwords and synonyms were lowercased; diacritics removed. In rare cases it produced duplicate records that were then removed, e.g. (OZH):

[3] Translated synsets are provided with glosses from original PWN synsets

Ex. 1. «Заброни́ровать»—см. брони́ровать. (Zabronírovat'—sm. bronírovat'). Reserve, book.

Ex. 2. «Забронирова́ть»—см. бронирова́ть. (Zabronirovát'—sm. bronirovát'). Armor, armour.

Additionally, two corpora were used in the study: Russian National Corpus (RNC, http://www.ruscorpora.ru) and Google Books Ngram Corpus (GBN, https://books.google.com/ngrams). RNC [29], first published in 2004, contains nowadays more than 192 mln tokens. In our study, we employed pre-processed RNC frequency lists (http://ruscorpora.ru/corpora-freq.html). GBN [19] contains year-by-year n-gram frequencies (up to 5-grams) from about 6% of all ever-published books in different languages. The Russian subcorpus of GBN contains about 103 billion tokens according to our calculations, which is much more than indicated by the authors—about 67 billion tokens. It could be explained by differences in token counting. Only unigrams that contain letters (and possibly hyphens) were taken into account in our work. It should be noted that there are words written in Latin alphabet in the Russian subcorpus. Both corpora word lists were lemmatized with *mystem* (https://tech.yandex.ru/mystem).

4. Analysis of lexical resources

4.1. Word lists analysis

Word lists of resources under consideration cover different parts of the Russian lexicon. The size of the word list itself is not sufficient to draw any conclusions. For example, WIKT contains about 35,000 proper nouns (about 18% of the whole volume). Moreover, authors of lexicographic resources treat derivative words, including gender-specific variants, in different ways. E.g. MAS contains separate entries for *«второклассник» (vtoroklassnik, «second-grade school boy»)* and *«второклассница» (vtoroklassnitsa, «second-grade school girl»)*, whereas BTS contains *«второклассник» ("second-grade school boy»)* as headword, and *«второклассница» ("second-grade school girl»)* as a variant.

Dictionaries' overlap seems to be a more suitable measure. Table 2 shows pairwise overlaps (in thousands) above the main diagonal and share of the overlap in the whole dictionary for the smaller resource in the pair below the main diagonal.

It was anticipated that there is a high degree of overlapping (80–90% in average) between the classical explanatory dictionaries (Table 2). The most intersecting dictionaries are MAS and BTS that share the same initial data sources [14]: 90.5% of MAS word list was included in BTS. Also note that WIKT word list includes many words from the classical explanatory dictionaries. This finding could also be explained by the large number of WIKT entries. RuThes-lite (RUT) and Russian Wordnet (RWN) contain a lot of multiword expressions (about 50% and 30% of the word list, respectively, see Table 1), it leads to smaller overlaps with other explanatory dictionaries.

Table 2. Overlaps between dictionary word lists

	BTS	EFR	MAS	OZH	RUT	RWN	USH	WIKT	ZLZ
BTS		85.8	73.8	38.3	39.2	18.8	63.3	80.1	72.5
EFR	0.831		74.2	38.1	38.5	19.4	70.0	89.3	80.5
MAS	0.905	0.909		36.3	36.0	17.6	61.2	66.6	66.8
OZH	0.951	0.945	0.901		22.8	13.2	35.0	36.8	37.3
RUT	0.406	0.398	0.441	0.567		14.2	31.9	41.6	36.7
RWN	0.611	0.628	0.571	0.428	0.461		17.4	20.1	18.9
USH	0.727	0.803	0.750	0.868	0.366	0.564		62.1	68.6
WIKT	0.776	0.722	0.817	0.912	0.430	0.653	0.713		79.4
ZLZ	0.776	0.862	0.819	0.926	0.393	0.612	0.787	0.850	

While Table 2 quantifies the overlap between dictionaries, Fig. 1 depicts the number of unique words in the resources (words that are presented only in one dictionary). In order to make this comparison more fair for traditional dictionaries we filtered out proper names from WIKT and multiword expressions from all resources.

The analysis with and without proper names and MWEs resulted in several findings:

1) Proper names and MWEs constitute one third of all lexical units.
2) In RWN and RUT there are 9 and 50 thousand of unique headwords respectively, but there are only 1 and 6 thousand after the removal of MWEs (Fig. 1).
3) The filtering of proper names and MWEs in WIKT reduced to half the number of unique headwords—from 87 to 47 thousand (Fig. 1).
4) 5 traditional dictionaries contain 731 MWEs out of 60,000 lexical units.

Both charts illustrate succession in creation of Russian explanatory dictionaries (BTS, MAS, USH, OZH, and EFR): there are 107,800 words occurring in at least two of these dictionaries. In this regard, Russian Wiktionary (WIKT), Russian WordNet (RWN), and RuThes-lite (RUT) contain almost twice as less crossings (i.e. words that are represented in at least two dictionaries out of three): 46,200[4] words. 59,500 words (see Fig. 1) correspond to a union of words from traditional dictionaries not included in any other Russian dictionary. These data can be useful for creating new explanatory dictionaries and for the further developing of WIKT and RUT.

[4] This number is not presented on Fig. 1.

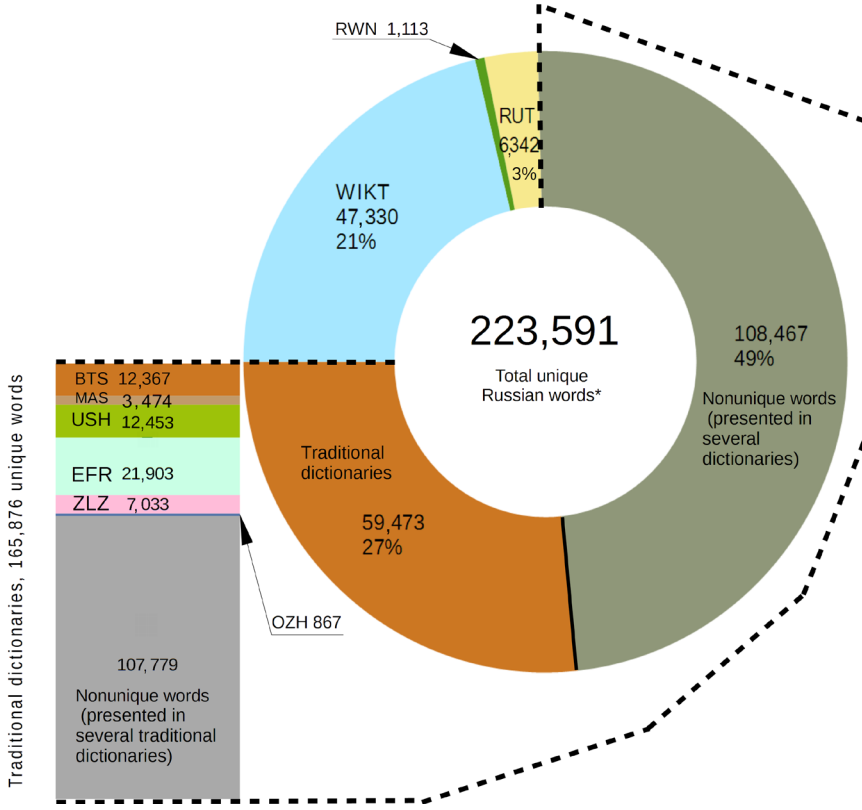


Fig. 1. Number of unique words in traditional dictionaries (vertical stripe) and in all dictionaries (pie chart)

*) Three smallest values—ABR (450 unique words), BAB (290) and EVG (130)—are not depicted in the pie chart, but accounted for in the total.

4.2. Corpora coverage

On the next stage of our study we quantified coverage of RNC and GBN with respective dictionary headwords. We employed two approaches to measure corpora coverage: 1) overlap of dictionary word lists and top-frequency lists extracted from corpora and 2) direct coverage of corpora (excluding stopwords).

The first approach simply measures intersection of word lists, the second one takes word frequencies in corpora into account. After lemmatization of the RNC frequency list around 263,000 unique terms remain; the number of unique lemmas in GBN corpus is more than 1.7 mln. The latter large number is partially due to a high level of misprints and systematic OCR errors [19]. Both corpora contain also a significant portion of proper names. We considered only top-100k most frequent lemmas for each corpus (the overlap of these two lists is 68,000 terms). Fig. 2 shows the presence

of the most frequent words from each of the corpora in dictionaries. For example, from the 1,000 most frequent RNC words 95.3% of them are presented in WIKT (i.e. 47 words are absent).

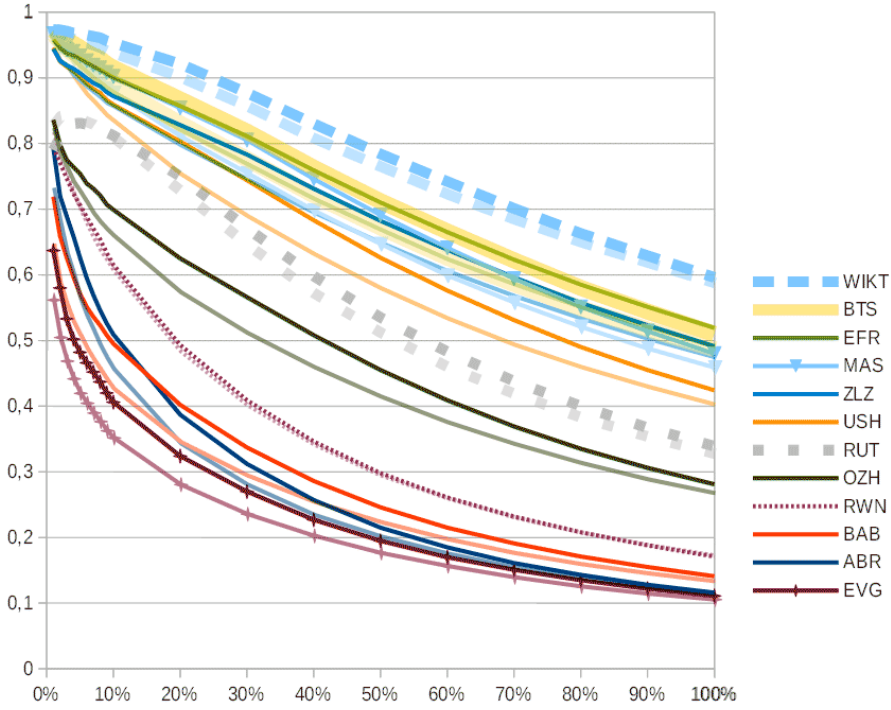


Fig. 2. Coverage of top-100k terms from RNC (solid lines) and GBN (softer lines)

On the right-side pane in Fig. 2 the dictionaries are listed in descending order of coverage for both corpora wordlists. The Figure clearly shows three groups of resources: 1) modern large dictionaries with good coverage (WIKT, BTS, EFR, and MAS); 2) borderline dictionaries (USH, RUT, and OZH); 3) synonym dictionaries with a lower coverage (RWN, BAB, ABR, and EVG). It is important to note that the dictionaries' ranks are the same for both corpora. This allows us to be more confident in generalizing the conclusions obtained from the data of either of two corpora.

The second approach accounts for all words presented in the corpora dataset (see Sec. 2) along with their frequencies except for stopwords that account for 34.5% and 28.8% tokens in RNC and GBN, respectively. To make the comparison fair for wordnets that typically do not contain functional words, we excluded stopwords from calculation. Corpora coverage with the dictionary words lists is shown in Fig. 3.

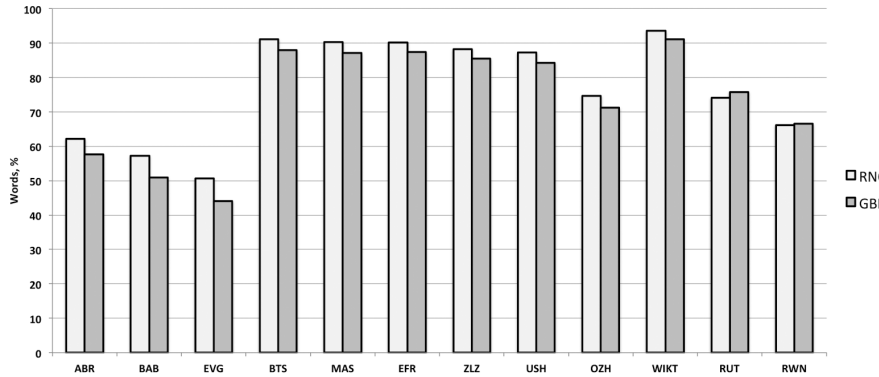


Fig. 3. RNC and GBN coverage

4.3. Analysis of modern lexicon coverage

As the results in the previous section show, all lexical resources cover core lexis quite well. We used GBN to evaluate how dictionaries under consideration reflect neologisms. To this end, we chose two 20-year intervals: 1970–1989 and 1990–2009, and selected lemmas that appeared at least in one thousand books in each time period. Then we ordered them by descending ratios of frequencies in the newer / older subcorpora and took top-2k words. The list contains many proper names, OCR errors, spelling variants and results of incorrect lemmatization. However, according to our manual evaluation, about a half of the list can be regarded as good 'headword candidates'. Fig. 4 shows how many neologisms from the 2,000 are presented in the dictionaries (Fig. 4 shows only dictionaries covering at least 50 lemmas). It is interesting to note that the attempt to create a list of obsolete words in the same simple way (by ordering the list by ascending frequency ratios) did not succeed: all top words were OCR errors or typos.

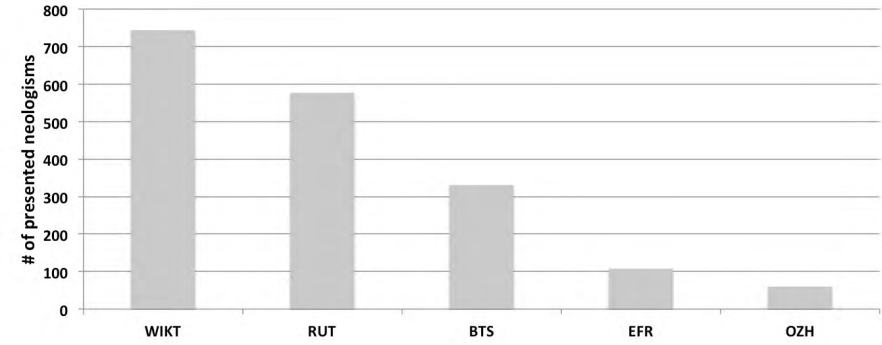


Fig. 4. Neologisms presented in dictionaries (from the list of 2,000 words)

4.4. Synonymy analysis

Synonymic resources are presented by three printed dictionaries of synonyms (ABR, EVG, and BAB), two thesauri (RWN and RUT), and WIKT. The latter combines properties of explanatory dictionary and dictionary of synonyms. All these dictionaries form their synsets/concepts uniquely, except WIKT, whose synsets are attached to respective headwords and are not necessarily coordinated, cf.:

Ex. 1. ***«Собака» (sobaka, «dog»)**, «пёс» (pyos, «dog»), «псина» (psina, «dog»), «друг человека» (drug cheloveka, «friend of a man»), «четвероногий друг» (chetveronogiy drug, «four-legged friend»).*

Ex. 2. ***«Пёс» (pyos, «dog»)**, «собака» (sobaka, «dog»), «кобель» (kobel', «male dog»).*

These two synsets, created from two different entries (for headwords «*Собака*» and «*Пёс*» respectively), will be treated as reflecting different meanings. However, synsets represented by the same set of words are considered as equal. Note that we did not consider one-word synsets[5] as well. For every synset we counted all pairs formed from its words (i.e., a synset consisting of 4 words forms $4 * 3/2 = 6$ synonym pairs). Table 3 summarizes data for six dictionaries[6].

Table 3. Quantitative characteristics of synsets from different dictionaries

	ABR	BAB	EVG	RUT	RWN	WIKT
Total # of synsets, thousands	7.5	4.9	5.4	22.7	11.0	33.0
Average synset size, words	6.8	6.0	3.9	4.9	2.2	2.9
Total # of synonym pairs, thousands	125.9	107.5	45.7	378.1	15.0	121.1

For synsets comparison we made an assumption that any pair of synset's terms defines roughly the meaning of the synset. This is quite a strong assumption that is often violated. Two following examples (from BAB) show that even when one synset is a subset of another synset they still may have different meanings:

Ex. 1. *«Начинающий» (Nachinayushchiy, «beginner»), «дебютант» (debyutant, «debiutant»), «новенький» (noven'kiy, «newcomer»), «новичок» (novichok, «novice»)—тот, кто впервые выступает на сцене, участвует в соревнованиях; делает первые шаги на каком-либо публичном поприще (a person who performs on stage for the first time, participates in competition; makes his/her first steps in any public arena).*

Ex. 2. *«Начинающий» (Nachinayushchiy, "beginner"), «дебютант» (debyutant, "debutant")—недавно приступивший к какому-либо роду деятельности (о человеке, группе лиц и т.п.) (a person who recently begun any kind of activity (about a person, group of individuals, etc.)).*

---

[5] E.g. RWN contains 14,000 one-word synsets.

[6] Dictionaries of synonyms cannot include synsets consisting of just one word by design, yet thesauri (RWN, RUT) can. So we considered only 2+ word synsets.

So we calculated the number of synset pairs between dictionaries, that Jaccard similarity coefficient is no less that 0.5 (Table 4). We analyzed synsets consisting of two or more words; so all "overlapped" pairs have two or more words in common. Note that this method takes both within-dictionary (main diagonal) and intra-dictionary overlaps into account.

Table. 4. Synset overlapping

	ABR	BAB	EVG	RUT	RWN	WIKT
ABR	20,370	400	590	90	410	1,290
BAB		880	2,100	410	840	2,680
EVG			830	440	1,210	4,080
RUT				1,380	350	1,290
RWN					1,810	4,390
WIKT						12,620

As we can see from Table 4, EVG has greatly influenced the later Russian dictionaries of synonyms.

4.5. Quantitative analysis of definitions

It is natural to expect that comprehensive dictionaries differ not only by the size of their word lists, but also by a number of definitions in them. In order to compare the resources in this regard, we analyzed seven dictionaries out of 12 discussed in the paper. We treated shades of meaning (usually separated by a double vertical line) as separate definitions.

However because of ambiguous formatting of electronic versions of explanatory dictionaries we had, sometimes it was impossible to get all definitions for an entry. This is particularly true for «noticeable shift in meaning» [8], labeled by a single vertical line. So it could lead to slight inaccuracies in measurements, caused by detecting not all meanings for headwords. Nevertheless we suppose that it did not significantly affect the result of our experiments.

Table 5 shows the quantitative characteristics of definitions from dictionaries under consideration.

Table 5. Quantitative characteristics of definitions in dictionaries

	USH	MAS	BTS	EFR	WIKT	BAB	RUT
Unique definitions, thousands	110.4	121.5	97.3	155.6	81.0	4.1	10.4
Avg. definition length, words	5.43	5.37	5.10	5.51	6.87	10.19	7.21
# of words, thousands	72.2	77.2	67.6	94.7	45.4	4.1	28.4
Avg. # of definitions per entry	1.56 (1.47)	1.64 (1.63)	1.50 (1.47)	1.72 (1.62)	1.89 (0.83)	1.00 (1.00)	1.25 (0.57)

When calculating characteristics in Table 5 we considered only headwords presented in at least two resources and having at least one definition. Values corresponding to the whole set of entries (i.e. without filtration) are presented in parentheses (see the last row in Table 5).

One can see from Table 5 that the average definition length and the average number of definitions are similar for all traditional dictionaries, yet the same characteristics of electronic resources differ significantly. E.g., WIKT and RUT contain many entries without definitions at all, which obviously lowers the average number of definitions per entry. WIKT word list includes proper names that usually have only one meaning and do not occur in other dictionaries. The average number of word meanings (definitions) for an entry in WIKT is 1.89 essentially depends on part-of-speech. By 2011, the average number of definitions for verbs was 2.5, whereas for nouns, adjectives and adverbs the value laid in range 1.5–1.7 [30].

The distribution of entries by the number of definitions is shown in Fig. 5.

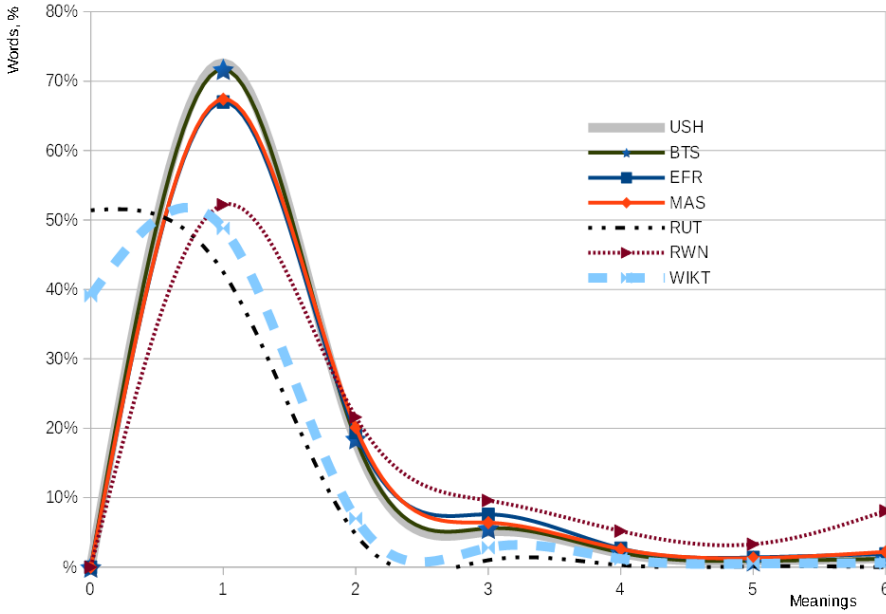


Fig. 5. Distribution of entries by the number of meanings (definitions)

Fig. 5 allows for interesting observations. Firstly, the distributions for USH and BTS almost coincide (both dictionaries have the largest share of monosemantic headwords). Secondly, the number of single-meaning words divides resources into two classes: academic (BTS, MAS, EFR and USH—a share of unambiguous words is about 70%) and other dictionaries (RUT, RWN and WIKT, drawn by dashed line, where unambiguous words comprise less than a half.

Note that we did not remove entries with zero definitions. Presence of some word means that a dictionary reflects it, but the quality of this reflection may vary and depend among other on definitions (some quantitative characteristics of definitions were discussed earlier).

4.6. Analysis of textual similarity of definitions

On the next stage of our study we compared textual similarity of same-word definitions in different dictionaries (note that we did not try to align the meanings of definitions). To this end we employed Monge-Elkan string similarity measure that combines word- and character-level similarity, demonstrates high performance and good balance between precision and recall [15, 5, 26].

Monge-Elkan similarity is not symmetrical, so we used the year of the first dictionary edition for selection of the direction of comparison (see Table 1). This direction reflects how definitions in newer resources resemble their predecessors' ones. In our study we used *DKPro Similarity* implementation of Monge-Elkan method [2].

Dictionaries contain a large number of 2–3 word definitions, which can skew similarity measurements, since such definitions are rather "standard" and occur in many dictionaries. So we filtered out such definitions, which resulted in exclusion of 176,000 lexical units. Typical examples of excluded definitions are:

- A widely used synonym;
- Ex. *«Помешкаться»—задержаться. (Pomeshkatsya—zaderzhatsya). Delay, linger.*
- lists of synonyms;
- Ex. *«Утопист»—мечтатель, фантазер. (Utopist—mechtatel', fantazer). Dreamer, visionary.*
- A gloss without examples.
- Ex. *«Манка»—манная крупа. (Manka—mannaya krupa). Semolina.*

As similar we considered definitions with similarity value above 0.9. Fig. 6 depicts textual similarity of definitions in different resources as a graph: vertices are dictionaries; edge thickness is proportional to the number of similar definitions in a pair of resources; borrowings from an older to a newer dictionary are displayed clockwise; numbers reflect the percent of borrowings in recipient dictionary.

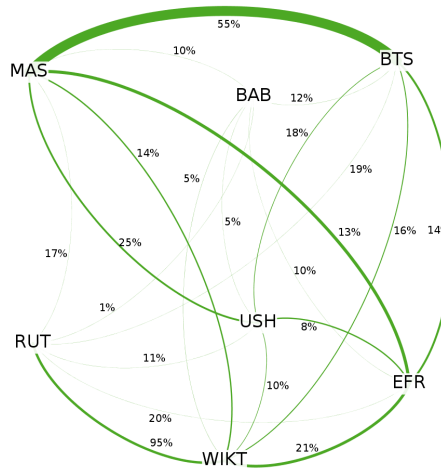


Fig. 6. Graph of textual similarity of definitions

5. Conclusion

Our results let us make the following conclusions.

1. Overlaps between dictionary word lists. A developed tradition and succession of different dictionary creation projects explain significant overlaps between word lists of traditional dictionaries. USH dictionary stands out in this regard, which could be explained by the fact that USH project is has not been developing anymore. A low overlap between RWN and other dictionaries, on the contrary, indirectly confirms the idea that a straightforward translation of a thesaurus into another language significantly reduces lexicon.

2. Number of unique words in dictionaries. As we found out, there are relatively few unique words and phrases (i.e. contained in only one dictionary). This fact is partly due to the choice of representation of derivatives—as a separate headword or inside an entry. At the same time in these dictionaries (even in EFR) there is a significant lack of multiword expressions, which are presented in electronic resources much better. In addition, a large number of unique terms in WIKT can be explained by the fact that its word list includes proper names (35,000 words out of 193,500).

3. Corpora coverage. Judging by the share of unique words, one would assume that the traditional dictionaries do not have good corpora coverage. However that is not true—especially with respect to BTS, EFR, and MAS. A noticeable "deficiency" of dictionaries of synonyms is quite clear and expected. The obtained results can give a raw estimate of Russian lemmas that are involved in synonymy relationships—about 60–70%.

4. Quantitative analysis of definitions. The share of monosemantic words, contained in traditional dictionaries, was significantly higher than in the electronic resources. This fact indicates the orientation of the latter towards actual word usage and a tendency to represent specific meanings.

5. Analysis of modern lexicon coverage. Finally, a comparison of dictionaries by presence of neologisms shows a great potential of modern electronic resources that can be dynamically modified. It does not mean that traditional dictionaries are obsolete. The lag from changes in a language gives an opportunity to reflect in the dictionary not just random, but established language phenomena: words, meanings, variations, etc.

The current situation in modern Russian lexicography reflects the transition period from traditional printed editions to large-scale projects based on large corpora and crowdsourcing. Traditional dictionaries based on manual sampling and data processing are regarded as high-quality sources, yet they are clearly behind the resources like WIKT, considering their volume and coverage of modern lexicon. At the same time, the specifics of electronic projects are often criticized for their quality.

We expect that our findings will be helpful for lexicographic practice—no matter what form will be chosen by dictionary authors.

Acknowledgment

Pavel Braslavski's and Mikhail Mukhin's contribution to the study was supported through grant #13-04-12020 "New Open Electronic Thesaurus for Russian" from the Russian Foundation for the Humanities. Mikhail Mukhin's work is also supported through Ural Federal University Competitiveness Enhancement Program # 02.A03.21.0006. Some parts of the research of Andrew Krizhanovsky are carried out in the project supported by grant # 15-04-12006 from the Russian Foundation for the Humanities, and the project "Veps corpus: computer morphological base development" of the basic research program of the Literature and language section of Department of history and philology RAS "Language and information technology" 2015–2017. We also thank Natalia Loukachevich for granting us access to RuThes-lite data.

References

1. *Abramov N.* Russian dictionary of synonyms and similar expressions on sense, M.: Russian dictionaries, 1999.
2. *Bär D., Zesch T., Gurevych I.* DKPro Similarity: An Open Source Framework for Text Similarity //ACL (Conference System Demonstrations).—2013.—P. 121–126.
3. *Braslavski, P., Mukhin, M. Y., Lyashevskaya, O. N., Bonch-Osmolovskaya, A. A., Krizhanovsky, A. A., & Egorov, P.* (2013). Yarn Begins. Proceedings of Dialog-2013.
4. *Braslavski, P., Ustalov, D., & Mukhin, M.* (2014). A spinning wheel for YARN: user interface for a crowdsourced thesaurus. In Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics, Gothenburg, Sweden (pp. 101–104).
5. *Cohen W., Ravikumar P., Fienberg S.* A comparison of string metrics for matching names and records //KDD Workshop on Data Cleaning and Object Consolidation.—2003.—T. 3.—C. 73–78.
6. Dictionary of synonyms, ed. A. Evgenyeva, L.: Nauka, 1975.
7. Dictionary of synonyms of the Russian Language, ed. L. Babenko. M.: AST: Astrel, 2011.
8. Dictionary of the Russian Language (Malyy akademicheskiy slovar'). Four volumes. RAS, Institute for Linguistic Studies, ed. A. P. Evgenyeva, M: Russkiy yazyk; Poligrafresursy, 1999.
9. Dictionary of the Russian Language (Tolkovyy slovar' russkogo yazyka). Four volumes, ed. D. N. Ushakov, editor. 1935–1940, State Publishing House of Foreign and National Dictionaries.
10. Explanatory Dictionary of Russian Language: 80,000 words and set phrases. Eds. S. I. Ozhegov and N. Yu. Shvedova. Moscow: Azbukovnik, 1999.
11. Fellbaum, Christiane. WordNet. Blackwell Publishing Ltd, 1998.
12. *Gel'fejn'bejn I. G., Goncharuk A. V., Lehel't V. P., Lipatov A. A., Shilo V. V.* (2003), Automatic translation of Wordnet in russian [Avtomaticheskij perevod semanticheskoj seti Wordnet na russkij jazyk], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2003"

[Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog 2003"], Protvino.
13. *Geyken, Alexander, and Lothar Lemnitzer.* «Using Google books unigrams to improve the update of large monolingual reference dictionaries.» In Proceedings of the 15th EURALEX International Congress, pp. 362–366. 2012.
14. Great Dictionary of the Russian Language (Bol'shoy tolkovyy slovar'), ed. S. A. Kuznetsov, St. Petersburg, Norint (1998).
15. *Jimenez S.* et al. Generalized Mongue-Elkan method for approximate text string comparison //Computational Linguistics and Intelligent Text Processing.—Springer Berlin Heidelberg, 2009.—C. 559–570.
16. *Kennedy, Alistair, and Stan Szpakowicz.* «Evaluating Roget's Thesauri.» In ACL, pp. 416–424. 2008.
17. *Kilgarriff, Adam, and Iztok Kosem.* «Corpus tools for lexicographers.» Electronic lexicography (2012): 31–56.
18. *Krizhanovsky A., Smirnov A.* An approach to automated construction of a general-purpose lexical ontology based on Wiktionary // Journal of Computer and Systems Sciences International, 2013, Vol. 52, No. 2, pp. 215–225.
19. *Yuri Lin, Jean-Baptiste Michel, Erez Lieberman Aiden, Jon Orwant, Will Brockman, Slav Petrov.* Syntactic annotations for the Google Books Ngram Corpus. Proc. of the ACL 2012 System Demonstrations, p. 169–174, July 10–10, 2012, Jeju Island, Korea
20. *Litkowski, Kenneth C.* Towards a meaning-full comparison of lexical resources // Association for Computational Linguistics SIGLEX Workshop. 1999.
21. *Loukachevitch N. V., Dobrov B. V., Chetviorkin I. I.* RuThes-lite, A publicly available version of thesaurus of russian language RuThes // Computational Linguistics and Intellectual Technologies: Conference "Dialogue", Issue 13 (20), pp. 340–349, Moscow, RGGU, 2014.
22. *Luk, Robert WP, and Venus MK Chan.* «A Quantitative Analysis of Word-Definition in a Machine-Readable Dictionary.» (1995).
23. *Michael Matuschek and Iryna Gurevych.* "High Performance Word Sense Alignment by Joint Modeling of Sense Distance and Gloss Similarity." In Proceedings of the the 25th International Conference on Computational Linguistics (COLING 2014), pp. 245–256.
24. *Meyer, Christian M., and Iryna Gurevych.* «Wiktionary: a new rival for expert-built lexicons? Exploring the possibilities of collaborative lexicography." Electronic Lexicography (2012): 259–291.
25. *Michiels, Archibald, and Jacques Noël,* Approaches to Thesaurus Production // In Proceedings of the 9th conference on Computational linguistics-Volume 1, pp. 227–232. Academia Praha, 1982.
26. *Monge A. E.* et al. The Field Matching Problem: Algorithms and Applications // KDD.—1996.—C. 267–270.
27. *Navigli, Roberto, and Simone Paolo Ponzetto.* "BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network." Artificial Intelligence 193 (2012): 217–250.
28. New dictionary of Russian: the sensible and word-formation (Novyy tolkovo-slovoobrazovatel'nyy slovar' russkogo yazyka), ed. T. F. Efremova, 1996.

29. *Plungyan, V. A., T. I. Reznikova, and D. V. Sichinava*. «The National Corpus of the Russian Language: general characteristics.» Nauchno-tekhnicheskaia informatsiia, ser. 2 (2005): 913.
30. *Smirnov A. V., Kruglov V. M., Krizhanovskiy A. A., Lugovaya N. B., Karpov A. A., Kipyatkova I. S.* A quantitative analysis of the lexicon in Russian WordNet and Wiktionaries (In Russian) // In Trudy SPIIRAN (St. Petersburg, 2012), Issue 23, pp. 231–253.
31. *Yoshida, Sho, Hiroaki Tsurumaru, and Tooru Hitaka*. «Man-assisted machine construction of a semantic dictionary for natural language processing.» In Proceedings of the 9th conference on Computational linguistics-Volume 1, pp. 419–424. Academia Praha, 1982.
32. *Zaliznyak A.* Russian Grammar Dictionary (Grammaticheskij slovar' russkogo jazyka). Moskva, 1977.

ГУЛЯЛ — НАГУЛЯЛ — НАГУЛЯЛСЯ. ЗАМЕТКИ О СТРУКТУРЕ ПРЕФИКСАЛЬНО-ПОСТФИКСАЛЬНЫХ ГЛАГОЛОВ

Киселева К. Л. (xenkis@mail.ru)
Институт русского языка РАН, Москва, Россия

Татевосов С. Г. (tatevosov@gmail.com)
МГУ и НИУ ВШЭ, Москва, Россия

Ключевые слова: деривационная морфология, циркумфиксы, структура события, деятельности, свершения

NOTES ON THE STRUCTURE OF CIRCUMFIXAL VERBS IN RUSSIAN

Kisseleva X. L. (xenkis@mail.ru)
Vinogradov Institute for Russian Languagee, Moscow, Russia

Tatevosov S. G. (tatevosov@gmail.com)
Lomonosov Moscow State University and
Higher School of Economics, Moscow, Russia

The paper argues for an analysis that reduces the derivation of non-compositional circumfixal verbs to a fully compositional combination of two pieces of morphology independently attested in Russian, a (resultative) prefix, and the reflexive morpheme *-sja*. Circumfical verbs are analyzed as involving the following steps of derivation. First, the activity event structure projected by a non-derived stem is augmented by the change-of-state component that turns it into an accomplishment even structure. Secondly, prefixation occurs that introduces the maximal degree of change along a relevant scale creating a transitive verb. Such a verb, however, is ill-formed since the change-of-state component has not been licensed via lexical insertion. To rescue the derivation, reflexivization is invoked, and the change-of-state subevent gets licensed through identification of its participant with the clausal subject. A wider theoretical implication of the analysis is that circumfixation, as a primitive type of affixation, is superfluous and is to be abandoned.

Key words: derivational morphology, circumfixation, event structure, activities, accomplishments

Цель этой работы[1] — сделать несколько шагов на пути избавления русской грамматики от таких способов словообразования, таких, как префиксально-суффиксальный, суффиксально-постфиксальный и префиксально-постфиксальный. Последний иллюстрируется примерами типа (1) из АГ80:

(1) *a. Артисты еще впоются,* ***выграются*** *в свою роль.*
b. ***Взбубнилась*** *злоба.*
c. Илья не ***выхворался****.*
d. При громадной усидчивости, трудоотдаче он может ***дописаться*** *до прозы, которая будет почти точной копией хорошей.*
e. Не слишком ли мы ***запраздновались?***
f. Все ***иззавидовались.***
g. — А ну встать! — Отставить! Уже ***навставались****.*
h. ***Опившись*** *квасу, Яков пошел на берег.*
i. Нинка ***обрыдалась*** *и обещала перемениться.*
j. ***Откукарекался****, — усмехнулся Игнат.*
k. Никак не могу я к ней ***подтанцеваться.***
*l. Стараюсь «****прииграться****» к нападающему, понять «язык» егодвижений.*
m. Он ***проорался****, и мы пошли к Ольге.*
o. У нас тут ползавода ***разгрипповалось****.*
p. Все ***слеглись*** *в одной палатке.*
q. Я прямо ***ухихикался.***

Такие глаголы образуются, согласно АГ80, присоединением сложного форманта (в других терминах — циркумфикса), состоящего из префикса и постфикса (*в-ся, на-ся, раз-ся* и т. д.). *На-ся*, например, имеет значение «действие, названное мотивирующим глаголом, совершить в достаточной степени или в избытке; дойти до состояния удовлетворения или пресыщения в результате длительного или интенсивного совершения этого действия».

Эмпирическая реальность циркумфиксов опирается на два факта: их значение идиоматично, а глаголы, в которых один из элементов циркумфикса отсутствует, не существуют в качестве самостоятельной лексической единицы[2].

---

[1] Исследование подержано грантом РФФИ № 14-06-00435. Авторы глубоко признательны четырем анонимным рецензентам «Диалога» за внимательное отношение к этой статье.

[2] Один из четырех анонимных рецензентов «Диалога» указывает: «Постулирование единой словообразовательной единицы циркумфикса *на-…-ся* обусловлена … 2) вхождением этого циркумфикса, как и ряда других, в языковую компетенцию говорящих на русском языке как самостоятельной единицы (ср. классическую работу И. А. Мельчука про „ты у меня допереключаешься“)». Это утверждение представляется несколько поспешным. Ни И. А. Мельчук, ни кто-либо другой, насколько нам известно, не владеет приемами, позволяющими превратить языковую компетенцию в объект, доступный для непосредственного наблюдения. Задача, которую И. А. Мельчук поставил перед собой в [Мельчук 1995], существенно скромнее: предложить описание соответствующей конструкции, удовлетворяющее критериям формальности, принятым в «Модели ‘Смысл — Текст’». В пределах этой статьи мы не имеем возможности обсуждать достоинства и недостатки этого описания.

Возьмем, например, *нагуляться* в (2). В русском языке у *нагуляться* нет переходного аналога *нагулять* (то есть глагола со значением «**довести** до состояния удовлетворения или пресыщения в результате длительного или интенсивного гуляния», (3a); о глаголе *нагулять* в случаях типа *нагулять аппетит* будет сказано ниже). Нет и глагола *гуляться*, (3b). Это как будто исключает деривационные цепочки в (4) и требует признать анализ в (5) единственно возможным.

(2) *Володя нагулялся.*

(3) *a. *Володя нагулял Феликса.*
*b. *Володя гулялся.*

(4) *a. Гулять → **на**-гулять → нагулять-**ся***
*b. Гулять → гулять-**ся** → **на**-гуляться*

(5) *Гулять → **на**-гулять-**ся***

Чем плоха «циркумфиксация» с теоретической точки зрения? Тем, что она увеличивает репертуар элементарных позиционных типов аффиксов, наращивает объем параметров межъязыкового варьирования и катастрофически усложняет задачу усвоения первого языка.

Если циркумфиксы типа *на-ся* непредставимы как сумма двух аффиксов, универсальная грамматика должна содержать циркумфиксацию как элементарную морфосинтаксическую операцию такого же уровня, как префиксация и суффиксация. Это немедленно порождает разветвленную языковую типологию (языки с суффиксами, языки с префиксами, языки с циркумфиксами, языки с суффиксами и циркумфиксами и т. д.) и предсказывает объемное межъязыковое варьирование. Усвоение языка обременяется необходимостью определить, где в огромном пространстве типологических возможностей располагается усваиваемый язык. Если универсальны только префиксация и суффиксация, типология сводится к четырем возможностям (что-то одно, что-то другое, и то и другое, ни то ни другое), а усвоение становится компактной задачей с понятным алгоритмом решения. Именно поэтому значительные достижения фонологии и морфологии последних десятилетий связаны именно с идеей сведения «неканонических» позиционных типов аффиксов (инфиксов, трансфиксов и т. п.) к двум элементарным операциям — префиксации и суффиксации.

Например, классический пример языка с инфиксацией, тагальский, при ближайшем рассмотрении превращается в язык с префиксацией (Prince, Smolensky 1993). Тагальский инфикс — это префикс, который может уточнить свою линейную позицию с целью построить слово с оптимальной слоговой структурой — с минимальным количеством закрытых слогов. Инфиксация в других языках — это в действительности префиксация, однако префиксация, определенная не в морфологических («присоединить префикс к основе»), а в просодических терминах (например, «присоединить префикс к первой стопе просодического слова») (McCarthy, Prince 1998).

Арабский язык — пример трансфиксации, в котором разрывные аффиксы помещаются внутрь разрывных корней. Однако, как мы знаем благодаря работам Дж. Маккарти (McCarthy 1979), в действительности в арабском языке, как и в любом другом, никаких разрывных морфем нет: этот эффект создается взаимодействием просодического шаблона словоформы с автономными («автосегментными») морфологическими уровнями — корневым и аффиксальным. Теоретические достижения такого рода открывают возможность избавить теорию языка от инфиксов и трансфиксов и выстроить более простой и более объяснительно адекватный инвентарь морфологических единиц.

Возможно ли так же избавиться от циркумфиксов русского типа? Ниже мы представим аргументы в пользу того, что значение циркумфиксов в действительности строго композиционально, а деривация осуществляется по сценарию в (4а). Мы выстроим изложение на материале глаголов так называемого интенсивно-результативного способа действия с «циркумфиксом» *на-ся*, предполагая, что наши выводы распространяются на другие глаголы из (1) без потери общности.

***

Чем различается семантика (6) и (7)?

(6) *Володя гулял.*

(7) *Володя нагулялся.*

(6) и (7) связаны отношением асимметричного следования: если верно, что Володя нагулялся, верно и то, что он гулял. Обратное неверно. Следовательно, (7) описывает более узкий класс ситуаций, чем (6). Прежде всего, *нагуляться* — глагол совершенного вида и в этом качестве предполагает предел, или кульминацию. В какой момент достигается кульминация? По мере развертывания описываемой ситуации меняется параметрическое свойство ее участника. Свойство может быть поименовано с помощью адъюнкта, как в (8), но может и оставаться недоспецифицированным, как в (7). В последнем случае его характер либо проясняется из контекста, либо так и остается неуточненным. Предложения типа (7) большинство носителей склонно интерпретировать как ‘гулял до исчерпания потребности в гулянии’ или ‘до состояния удовлетворенности’.

(8) *Я находился за день, нагулялся до изнеможения, наслушался безмолвия*
[Валентин Распутин. На родине (1999)]

Кульминация ситуации достигается в тот момент, когда свойство достигает определенной контекстно-зависимой степени. После кульминации наступает результирующее состояние, которое состоит в обладании этой степенью свойства. Эти простые наблюдения подводят нас к первому обобщению:

(9) Исходный глагол и циркумфиксальный дериват различаются структурой события.

Структура события — понятие, определяющее, как описываемая ситуация складывается из подсобытий, каковы дескриптивные свойства этих подсобытий и как актанты предиката соотносятся с подсобытийным компонентами (см. не утративший актуальности обзор имеющихся теорий в Levin, Rappaport Hovav 2005, а также разбор нескольких конкретных систем в Татевосов 2010). Для нас важны два типа событийных структур — структура деятельностей и структура свершений[3]. (10) и (11) показывают принимаемый нами событийно-структурный анализ непереходных деятельностей и свершений[4].

(10) $|| [V_{\text{деятельность}}] || = \lambda x.\lambda e.\text{process}(e) \wedge \text{arg}(x)(e) \wedge || V ||(e)$

(10) представляет собой отношение между индивидами и событиями. События, входящие в это отношение, обладают дескриптивными свойствами, определяемыми глаголом (|| V ||). Кроме того, любые деятельности удовлетворяют предикату process, который, в частности, гарантирует, что сложная событийная дескрипция λe. process(e) ∧ arg(x)(e) ∧ || V ||(e) является кумулятивной и неквантованной при любом x. Индивидные элементы, по которым пробегает переменная x, состоят с событиями в семантическом отношении arg. Для большинства глаголов деятельности arg — это семантическая роль агенса. Соответственно, глагольная группа *v*P предложений типа (6) имеет вид (11): это событийный предикат, обозначающий события, в которых Володя гуляет.

(11) $|| [_{vP}$ Володя гуля-] || = λe.process(e) ∧ agent(Volodja)(e) ∧ walk(e)

---

[3] Событийные структуры и акциональные классы не находятся во взаимнооднозначном соответствии. В первом приближении можно сказать, что в акциональном классе деятельностей (вендлеровские activities) содержатся только событийные структуры деятельностей. Предикаты, имеющие структуру свершений (accomplishments), попадают в акциональный класс свершений. Подробнее о том, как соотносятся акциональный класс и событийная структура см. Татевосов 2010.

[4] Здесь и далее мы используем экстенсиональное λ-исчисление с элементарными логическими типами e (индвиды), v (события), d (степени) и t (истинностные значения); производные типы рекурсивно определяются стандартным образом. (Термин «событие» здесь следует понимать как синоним термина «ситуация».) Наличие в этой системе событий делает ее вариантном так называемой событийной семантики (Davidson 1980, Parsons 1990 и обширная последующая литература). Некоторые аргументы в пользу придания событиям онтологического статуса суммируются в Татевосов 2013. Все прочие компоненты метаязыка и его интерпретация стандартны. В нем есть переменные, пробегающие по индивидам, событиям, степеням (соответственно, x, y, z, …, e, e', e''', …, d, d', d''', … и т.д.) и любым выражениями производных логических типов (например, переменная P соответствует одноместным предикатам любых типов); константы любых логических типов (например, walk — константа типа <v,t>) и т.д. Индивиды соотносятся с событиями с помощью семантических ролей (например, agent), имеющих логический тип отношений между индивидами и событиями <e, <v,t>>. Единственное правило композиции, используемое в этой статье, — применение функции к аргументу. Читателям, не знакомым с началами теоретико-модельной семантики, мы рекомендуем обратиться за дальнейшими подробностями к Heim, Kratzer 1998.

Событийная структура свершений устроена сложнее. Она состоит по меньшей мере из двух подсобытий: подсобытия деятельности, которое устроено так же, как (10), и подсобытия изменения состояния. Два подсобытия связаны отношением непосредственной каузации. Событийная структура свершений содержит по меньшей мере два индивидных актанта, по одному на каждое подсобытие.

(12) $|| [V_{\text{свершение}}] || = \lambda d.\lambda y.\lambda x.\lambda e.\text{process}(e) \wedge \text{arg}(x)(e) \wedge \exists e'[\text{increase}(G_V(y))(d)(e') \wedge \text{cause}(e')(e)]$

Самая существенная часть компонента, определяющего изменение состояния — отношение INCREASE между индивидами, событиями и степенями:

(13) || INCREASE(G(y))(d)(e′) || = 1 тогда и только тогда, когда степень, в которой индивид y обладает параметрическим свойством G, возрастает при осуществлении события e на степень d.

В (12) глагол определяет дескриптивные свойства не подсобытия деятельности, а подсобытия изменения состояния (переменная e′), причем делает это опосредованно — путем задания параметрического свойства G, которое меняется при осуществлении ситуации. (В (12) G фигурирует как несвязанная переменная, значение которой присваивается в результате оценки переменных.) Таким способом мы реализуем известную дихотомию глаголов способа и глаголов результата в принятом нами варианте анализа деятельностей и свершений. Наконец, последний элемент (12) — отношение непосредственной каузации, в которое вступают подсобытия деятельности и изменения состояния.

Прежде чем сформулировать семантику *нагуляться* в (7), рассмотрим более простой случай: глаголы свершений типа *высушить*. Глагольная группа в предложении типа *Володя высушил белье* получает анализ в (14):

(14) $|| [\text{Володя высуши- белье}] || = \lambda e.\text{process}(e) \wedge \text{agent}(V.)(e) \wedge \exists e'[\text{increase}(\text{dry}(\text{linen}))(d_{max})(e') \wedge \text{cause}(e')(e)]$

(14) обозначает множество процессов e, в которых агенсом выступает Володя. Каждый такой процесс вызывает к жизни событие e′, в котором сухость белья (dry(linen)) возрастает в максимальной степени d_{max}.

В этой статье мы предполагаем в порядке упрощения, что максимальная степень изменения параметрического свойства — это семантический вклад префикса. Соответственно, мы рассматриваем префикс как функцию, которая применяется к обозначаемому глагольной основой отношению между двумя индивидами, степенями и событиями. Префикс присваивает степенной переменной максимальную (по отношению к данному индивиду) степень изменения по шкале, соответствующей обозначаемому глаголом параметрическому свойству[5].

[5] Более полный анализ префиксов должен учесть то, что префиксы определяют результирующее состояние.

(15) || PRF || = λS.ly.λx.λe.S(y)(x)(max{d | ∃x∃e′ [S(x)(y)(d)(e′)]})(e) = λS.ly. λx.λe.S(y)(x)(d_{max})(e)

Для основы глагола *сушить* параметрическое свойство — ‘сухой’, а максимальная степень изменения — та, в которой достигается нулевая степень влажности.

Поскольку событийная структура свершений содержит в качестве аргумента степенную переменную, присоединение префикса делается необходимым для заполнения этой аргументной позиции. С интуитивной точки зрения это означает следующее: свершения описывают изменение и нуждаются в выражении, которое задает степень этого изменения. Соответственно, за вычетом нескольких специальных случаев, в русском языке любая структура свершений является префиксальной, и любая префиксальная конфигурация имеет структуру свершений.

Интерпретация префикса в общем случае зависит от двух параметров (в порядке упрощения это не показано в (15)): от конкретной лексической основы и от типа шкалы, представленного в конкретной событийной структуре. Например, с основой *суши-* префикс *вы-* опирается на минимальное значение на шкале влажности, лексически заданной для этой основы, а с основой *писа-* — на шкале, указывающей на полноту соответствия копии и оригинала: для *выписать цитату из книги* кульминация наступает, когда создание копии цитаты завершается. Это широко известный идиоматический аспект семантики префиксов (по крайней мере, так называемых лексических префиксов).

***

Перейдем к (7). Мы предполагаем для (7) следующий деривационный сценарий:

(16) а. Событийная структура деятельностей преобразуется в структуру свершений.
б. Структура свершений присоединяет префикс.
в. Образовавшаяся конфигурация содержит два референциально независимых индивидных аргумента и в этом качестве оказывается некорректно построенной.
г. Деривация восстанавливает корректность посредством рефлексивизации, отождествляющей две аргументные позиции.

Рассмотрим эти шаги по порядку. Исходный пункт деривации — структура деятельностей в (10). Глаголы типа *нагуляться*, как мы видели, обозначают изменение параметрического свойства, которого в (10) нет. Мы предполагаем, что этот шаг обеспечивается фонологически не выраженной операцией $SHIFT_{Д→С}$, которая создает из структуры деятельностей структуру свершений посредством приращения компонента изменения состояния.

(17) $SHIFT_{Д→С}$ (|| гуля- ||) = λd.λy.λx.λe.process(e) ∧ arg(x)(e) ∧ walk(e) ∧ ∃e′[increase(G(y))(d)(e′) ∧ cause(e′)(e)]

(17), как и любая структура свершений, содержит аргументную позицию для степени изменения параметрического свойства G, а значит, требует присоединения префикса. После этого образуется отношение в (18).

(18) || на- [$SHIFT_{Д \to С}$ (|| гуля- ||)] || = λy.λx.λe.process(e) ∧ arg(x)(e) ∧ walk(e) ∧ ∃e′[increase(G(y))($d_{МАХ}$)(e′) ∧ cause(e′)(e)]

(18) — отношение между двумя индивидами и событиями, которое обладает следующими свойствами. Индивид, соответствующий переменной x, гуляет. В результате у другого индивида у степень обладания некоторым контекстно заданным свойством прирастает до максимальной. Это семантика глагола *нагулять*.

По поводу этого глагола возникают два наблюдения. Во-первых, предложения типа (19) неграмматичны.

(19) **Володя нагулял Феликса (до изнеможения).*

В (19) мы предполагаем значение, идентичное значению *нагуляться*, но с той разницей, что изменение параметрического свойств происходят не с тем, кто гуляет, а с некоторым другим актантом. Меняющееся свойство, соответственно, — это один из параметров, характеризующих внутреннее состояние индивида, например, ‘утомленность’, ‘удовлетворенность’ и т. п. Как видно из (19), такую семантику *нагулять* иметь не может.

Неграмматичесность (19), однако, не означает, что глагол *нагулять* невозможен вовсе. В этом состоит второе наблюдение. *Нагулять* допускается в предложениях типа (20), где прямым дополнением выступает ИГ *аппетит*.

(20) *Володя нагулял аппетит.*

Среди немногочисленных вхождений глагола *нагулять* в НКРЯ более одного раза встречаются следующие прямые дополнения:

(21) *аппетит, жир(ок), брюхо, дитя, щетина, тело, сон, авторитет*

Как видно из (21), диапазон возможностей достаточно широк, и более того, основа *гуля-* задействована в разных значениях (ср. *нагулять ребенка, нагулять авторитет и нагулять жирок*). Все случаи объединяет общее свойство: объект создается при осуществлении (того или иного варианта) ситуации гуляния. Соответственно, *нагулять* в этих случаях — это глагол созидания, а параметрическое свойство G в (20) принимает облик ‘быть созданным’.

Это, однако, не все. Критически важный факт о структуре и семантике предложений типа (20) состоит в том, что подлежащее и прямое дополнение должны быть связаны отношением обладания (в широком смысле). Аппетит в (20) — это Володин аппетит, но никак не чей-то еще. Реализовав посессора при прямом дополнении в качестве дативного адъюнкта, мы получаем возможность сделать этот факт явным:

(22) *Володя нагулял себе аппетит.*

(23) **Володя нагулял Феликсу аппетит.*

Ограничение в (22)–(23) не вытекает из семантики в (18), однако оно помогает прояснить общую картину. Глагол *нагулять* с семантическим представлением в (18) возможен только как промежуточный шаг деривации, после которого мы должны направиться двумя путями: либо построить глагол *нагуляться*, либо создать глагол *нагулять,* у которого два актанта находятся в отношении обладатель-обладаемое. Эти этапы деривации показаны соответственно в (24)–(25) и (26)–(27):

(24) || -ся- || = λS.λx.λe.S(x)(x)(e)

(25) || [-ся [на- [$\text{SHIFT}_{\text{Д}\to\text{С}}$ (|| гуля- ||)]]] || = λx.λe.process(e) ∧ arg(x)(e) ∧ walk(e) ∧ ∃e′[increase(G(x))(d)(e′) ∧ cause(e′)(e)]

В (25) в результате рефлексивизации создается отношение между индивидами и событиями такое, что индивид одновременно выступает агенсом гуляния и объектом изменения параметрического свойства, которое в результате гуляния достигает максимума. Это в точности семантика глагола *нагуляться*.

(26) || *poss* || = λS.λy.λx.λe.S(y)(x)(e) ∧ poss(y)(x)

(27) || [*poss* [на- [$\text{SHIFT}_{\text{Д}\to\text{С}}$ (|| гуля- ||)]]] || = λy.λx.λe.process(e) ∧ arg(x)(e) ∧ walk(e) ∧ ∃e′[increase(G(y))(d)(e′) ∧ cause(e′)(e) ∧ poss(y)(x)]

В (26)–(27) описана посессивизация, которая сохраняет оба индивидных аргумента (18) нетронутыми, но наводит на них дополнительное ограничение: внешний аргумент должен быть посессором внутреннего.

В этом месте возникает два взаимосвязанных вопроса: что общего между рефлексивизцией и посессивизацией и почему та или другая необходимы для спасения (18) от деривационного краха? Ответ на первый вопрос напрашивается: и то и другое — способы определения одной сущности через другую. Рефлексивизация — наиболее радикальный способ, когда сущности делаются референциально идентичны. Посессивизация приводит к похожему эффекту: референция объектной именной группы становится зависимой от референции другой именной группы (или приводит к связыванию переменной квантором, если в позиции подлежащего — кванторная именная группа). И то и другое можно рассматривать как разновидности одной процедуры — определения участника одного подсобытия через участника другого. Назовем эту процедуру **соотнесением подсобытий через участника**, или Р-соотнесением.

Почему Р-соотнесение с необходимостью применяется к отношению в (18)? Ответ на этот вопрос начинает проясняться с помощью следующего наблюдения:

(28) Р-соотнесение происходит ровно в тех случаях, когда лексическая основа не в состоянии определить дескриптивные свойства подсобытия изменения состояния.

У основ типа *сушить* характер изменения состояния определен лексическим значением основы: это изменение по шкале влажности. У *нагуляться* и *нагулять аппетит* семантика *гуля-* определяет характер подсобытия деятельности — это гуляние. Однако она ничего не сообщает о характере изменения состояния. Как мы отметили выше, тип параметрического свойства, которое изменяется в этих случаях, частично ограничивается префиксом и частично определяется контекстом (в технических терминах — с помощью оценки переменной G). Тем не менее, никаких лексических единиц, непосредственно внедряющих в семантическое представление информацию о дескриптивных свойствах подсобытия изменения состояния, в этом случае нет. (18) — дефектная структура свершений.

Мы предполагаем, что именно недоопределенность подсобытия изменения состояния приводит к запуску процедуры Р-соотнесения. Подсобытия с недоопределенными дескриптивными свойствами доопределяются через соотнесение их участников с участниками другого подсобытия.

Какого типа данные позволили бы нам убедиться, что это предположение верно? В этом месте на помощь приходит следующее рассуждение. Если конфигурация в (18) плоха недоопределенностью подсобытия изменения состояния и если спасение деривации от краха происходит с помощью Р-соотнесения, мы ожидаем, что в тех случаях, когда подсобытие определено в достаточной степени, Р-соотнесения не происходит. Соответственно, мы ожидаем, что если основы типа *гуля-* возможны в переходных конфигурациях с референциально независимым прямым дополнением, то подсобытие изменения состояния обязано быть лексически идентифицировано. Поскольку единственный доступный лексический идентификатор — это сама основа *гуля-*, мы ожидаем, что в этих случаях она будет модифицировать не подсобытие деятельности, а подсобытие изменения состояния. Судя по всему, ожидание подтверждается.

Основы типа *гуля-* действительно допускаются в переходных конфигурациях с референциально независимым прямым дополнением:

(29) *Володя выгулял* ^{ОК}*свою*|| ^{ОК}*его собаку.*

В (30) участником ситуации гуляния выступает прямое дополнение, но не подлежащее:

(30) *Володя выгулял собаку.*
→ Собака гуляла.
*→ Володя гулял.

Чтобы (29) правильно описывало мироздание, требуется, чтобы гуляла собака. Чтобы гулял Володя, совершенно не требуется. (29) истинно, даже если Володя в течение всей ситуации неподвижно сидел на скамейке. Это показывает,

что предикат *гуля-* интегрируется в событийную структуру (29) принципиально иначе, чем в *нагуляться* или *нагулять аппетит*. В (29) он модифицирует подсобытие изменения состояния — то, в котором собака приобретает определенную степень выгулянности. Семантическое представление (29) показано в (31):

(31) || [вы- [$SHIFT_{Д \to С}$ (|| гуля- ||)]] || = λy.λx.λe.process(e) ∧ arg(x)(e) ∧
∃e′[walk(e) ∧ increase(G(y))($d_{МАХ}$)(e′) ∧ cause(e′)(e)]

Семантика в (31) верно предсказывает, что подсобытие изменения состояния — это одновременно событие гуляния, и его участником выступает внутренний аргумент предиката.

Глаголы типа *выгулять* в сопоставлении с глаголами типа *нагулять (аппетит)* и *нагуляться*, позволяют заключить, что высказанное выше предположение подтверждается. Необходимость Р-соотнесения наступает во всех и только в тех случаях, когда не происходит лексической идентификации подсобытия изменения состояния. Если это так, то свойства дериватов от основ типа *гуля-* становятся полностью предсказуемыми и регулируются единственным принципом в (32):

(32) Подсобытие изменения состояния должно быть лексически идентифицировано.

Эмпирическая реальность этого принципа находит независимые подтверждения в литературе по лексической декомпозиции, в первую очередь у Б. Левин и М. Раппапорт Ховав (Levin, Rappaport Hovav 1995, Rappaport Hovav, Levin 1998, Rappaport Hovav 2012). Р-соотнесение — это стратегия спасения деривации для тех случаев, когда (31) не выполняется. Похожая операция, называемая TPCONNECT, обсуждается С. Ротстин (Rothstein 2004) в связи с построением результативных конструкций в английском языке. Английские результативы, однако, в отличие от русских «циркумфиксальных» глаголов, не нуждаются в буквальном отождествлении участников двух подсобытий, поскольку подсобытие изменения состояния лексически определяется с помощью результативной ХР.

***

Подведем итог. Русские глаголы типа *нагуляться* не являются аргументом в пользу выделения циркумфисков как отдельного позиционного типа аффиксов. Данные, которые мы рассмотрели, полностью совместимы с анализом в (32):

(33) *Гулять* → ***на-****гулять* → *нагулять-****ся***

Мы предположили, что промежуточный шаг деривации — переходные глаголы типа *нагулять* с референциально независимым прямым дополнением — невозможны в силу того, что подсобытие изменения состояния остается в них лексически неспецифицированным. Присоединение *-ся*, однако, открывает путь к спасению: подсобытие доопределяется не через дескриптивные

свойства, а через идентификацию его участника с участником подсобытия деятельности. У *нагуляться* постфикс *-ся* выступает в прототипической ипостаси — как рефлексивный показатель, отождествляющий две аргументные позиции предиката. Если предложенный нами анализ верен, постулирование циркумфиксов лишается сколько-нибудь серьезных эмпирических оснований.

Литература

1. *Davidson D.* (1980), The logical form of action sentences, in Essays on Actions and Events, Clarendon Press, pp. 105–122.
2. *Hale K., Keyser S. J.* (2002), Prolegomenon to a theory of argument structure, MIT Press, Cambridge.
3. *Levin B., Rappaport Hovav M.* (1995), Unaccusativity. At the Syntax-Lexical Semantics Interface, MIT Press, Cambridge.
4. *Levin B., Rappaport Hovav M.* (2005), Argument Realization, Cambridge University Press, Cambridge.
5. *McCarthy J.* (1979), Formal Problems in Semitic Phonology and Morphology, PhD dissertation, MIT.
6. *McCarthy J., Prince A.* (1998), Prosodic morphology, in The handbook of morphology, Blackwell, Oxford, pp. 283–305.
7. *Parsons T.* (1990), Events in the Semantics of English: A Study of Subatomic Semantics, MIT Press, Cambridge.
8. *Prince A., Smolensky P.* (1993), Optimality Theory: Constraint interaction in generative grammar, Technical Report CU-CS-696-93, Department of Computer Science, University of Colorado at Boulder, and Technical Report TR-2, Rutgers Center for Cognitive Science, Rutgers University, New Brunswick.
9. *Rappaport Hovav M.* (2008), Lexicalized Meaning and the Internal Temporal Structure of Events, in Crosslinguistic and Theoretical Approaches to the Semantics of Aspect, John Benjamins, Amsterdam, pp. 13–42.
10. *Rappaport Hovav M., Levin B.* (1998), Building Verb Meaning, in The Projection of Arguments: Lexical and Compositional Factors, CSLI Publications, Stanford, pp. 97–134.
11. *Rothstein S.* (2004), Structuring Events: A Study in the Semantics of Lexical Aspect, Blackwell publishing, Malden.
12. Russian Grammar [Russkaja grammatica] (1980), Moscow.
13. *Tatevosov S.* (2010), Actionality in Grammar and Lexicon [Aktsionalnost' v leksike i grammatike.], D.Sc. dissertation, Moscow State University.
14. *Tatevosov S.* (2013), Event semantics as an empirical enterprise [Semantika sobytija kak empiricheskaja problema], Philosophy of Language and Formal Semantics [Filosofija jazyka i formalnaja semantika], Alpha, Moscow, pp. 8–42.

РЕДУКЦИЯ ГЛАСНОГО КАК ПОКАЗАТЕЛЬ ЕГО УДАРНОСТИ В СОВРЕМЕННОМ РУССКОМ ЛИТЕРАТУРНОМ ЯЗЫКЕ

Князев С. В. (svknia@gmail.com)

МГУ им. М. В. Ломоносова, НИУ ВШЭ, Москва, Россия

В докладе излагаются результаты экспериментально-фонетического исследования реализации первого гласного в двусложных составных единицах (типа *стоп-кран*) в сопоставлении с тем же гласным в простых словах (*стоптать*) в различных фразовых позициях. Полученные результаты свидетельствуют о том, что в двусложных составных единицах в слабой фразовой позиции при отсутствии ударения на первом слоге гласные среднего подъема [о] и [е] в первом предударном слоге после твердых согласных чередуются с [ъ] ([ə]), а не с гласным полного образования [а] или [ы], как в словах с одной основой. Произношение [ъ] ([ə]) в первом предударном слоге после твердых согласных является, таким образом, фонетическим показателем фонологической ударности гласного, которая в сильной фразовой позиции проявляется в полном отсутствии редукции. С фонетической точки зрения рассматриваемые единицы являются сочетанием двух фонетических слов. Полученные данные могут быть использованы при решении вопроса об орфографической адаптации заимствованных слов.

Ключевые слова: фонетика, ударение, редукция, ритмическая структура слова

VOWEL REDUCTION AS AN INDICATOR OF ITS STRESS IN STANDARD MODERN RUSSIAN

Knyazev S. V. (svknia@gmail.com)

Moscow State Lomonosov University, Higher School of Economics, Moscow, Russia

The phonetic unity of phonological word in Standard Modern Russian is governed to a large extend by phonological rules of vowel realization, which forbid the reduced vowel in a position of first pretonic syllable (with the exception of some clitical syntactic words — prepositions and particles). The paper deals with the instrumental study of phonetic markers of stress in compound (predominantly loan) disyllabic words in Standard Modern

Russian (e.g. *stop-krán* ‘emergency brake’) as compared with non-compound native words of the same phonological structure (*stoptát’* ‘tread down’). The paper states that compound disyllabic words under a phrase accent have both syllables stressed, stress being signalized by mid [o] // [e] vowels impossible in unstressed position in Standard Modern Russian. In non-compound native words unstressed [o] vowel in all types of phrase positions after “hard” consonants is displaced by [a] (unstressed [a] being about 10 percent longer than stressed [o]). Our data shows that in compound disyllabic words in a position with no tonal accent phonetically unstressed /o/ is realized by reduced [ə] (not standard [a]) vowel (being somewhat twice shorter than unstressed [a]). Thus, non-trivial [o] vowel reduction in compounds may serve as a phonetic cue of phonological stress which is shown up fully only under the tonal accent. Phonetically the units in question should be treated as a combination of two phonological words and phonetic data may be used as a ground for orthographic adaptation of loan word.

Key words: phonetics, word stress, vowel reduction, rhythmical structure of phonological word

Одной из наиболее ярких типологических особенностей современного русского литературного языка (далее — СРЛЯ) является ритмическая структура фонетического слова, описанная еще в конце XIX века при помощи формулы *…112́311…* [Potebnya 1866]. Необычность ее заключается в наличии так называемого просодического ядра, состоящего из ударного и первого предударного слогов, гласные которых отчетливо противопоставлены гласным во всех прочих слогах, в первую очередь, по длительности и своим спектральным характеристикам: только в них в подавляющем большинстве слов невозможны редуцированные гласные [ъ] и [ь][1], а мощность правила, запрещающего наличие редуцированных в 1-м предударном слоге такова, что его действие все в большей степени распространяется и на заударные гласные в слове перед ударным слогом следующего фонетического слова внутри синтагмы, которые в СРЛЯ характеризуются существенно большей длительностью, чем в позиции перед словом с безударным начальным слогом последующего слова [Knyazev 2007], [Grammatchikova, Knyazev 2014]. Исключения из этого правила немногочисленны и могут быть описаны как словарные свойства определенных лексем: таково, в частности, произношение [ъ] в позиции перед ударным гласным в служебных словах *да, вот, но, что, чтоб, хоть* (*хоть тресни, вот так, чтоб было, но там, да ладно*).

---

[1] Следует отметить, что речь в данном случае идет о так называемом «московском» варианте СРЛЯ. Результаты современных исследований в области диалектной фонетики позволяют утверждать, что в говорах русского языка реализация гласных в слове зачастую отличается от описываемой «формулой Потебни»: просодическое ядро слово может состоять только из одного ударного слога, а дополнительное выделение — при его наличии — могут получать второй предударный или начальный слоги [Vysotskiy 1973], [Paufoshima 1983], [Al’mukhamedova, Kul’sharipova 1980], [Knyazev, Urbanovich 2002]. Эти диалектные особенности, обусловленные особенностями просодической организации речи, в том числе, местом реализации фразового акцента в слове, плохо осознаются и с трудом контролируются говорящим, поэтому зачастую сохраняются в речи носителей литературного языка из соответствующих регионов [Grammatchikova, Knyazev, Lukyanova, Pozharitskaya 2013].

Существует, однако, и ещё одна группа слов, в которых, как показывают данные аудитивного анализа, произношение редуцированного гласного в позиции перед ударным является довольно частотным (как минимум, в слабых фразовых позициях) — это сложные слова, состоящие из двух основ, преимущественно заимствованных, первая из которых односложная и содержит гласный среднего подъема после твердого согласного, а вторая имеет ударение на первом слоге, например: *хэштег (хештег), флеш-бэк, флешмоб, фейспалм, поп музыка, стоп-кран, стоп-кадр, шорт-лист, шорт-трек, док файл, ворк шоп, дропбокс, ток шоу, нонстоп, бойфренд, бой-баба, пол-яблока и т.* п. Обычно в сильной фразовой позиции слова такого типа содержат два ударения: [сто́п кра́н], [по́б гру́ппа], [шо́рт тре́к], [бо́й фре́нт] (иногда, впрочем, это явление трактуется как наличие нередуцированного [о] в безударных слога[2]). В слабой фразовой позиции первое ударение может утрачиваться, при этом в первом слоге произносится не полный гласный, как это можно было ожидать внутри фонетического слова в соответствии с формулой Потебни ([шартл'и́ст], [стапкра́н], [палја́блъка]) и как это эксплицитно рекомендуется в [Kalenchuk, Kasatkin, Kasatkina 2012], например, для слов *полвека*, *полвзвода* и т. п. [Kalenchuk, Kasatkin, Kasatkina 2012: 592], а редуцированный ([шъртл'и́ст], [стъпкра́н], [пъља́блъка]).

Таким образом, тестируемая в данной работе гипотеза заключается в том, что фонетическая редукция гласного до [ъ] в первом предударном слоге может служить маркером его фонологической ударности (наличия ударения в иных просодических условиях в том же слове) в московском варианте современного русского литературного языка.

Для проверки этой гипотезы было проведено экспериментально-фонетическое исследование, в ходе которого, в первую очередь, сравнивалось произношение единиц *стоп-кран* (двухударной) и *стоптать* (с одним ударением). Выбор слова *стоптать* для сопоставления со словом *стоп-кран* был обусловлен их сходной фонетической структурой:

1) каждая из единиц содержит два слога,
2) второй из которых ударный [а́],
3) первый слог закрытый,
4) однотипный консонантный контекст гласного (между сочетанием [ст] и двумя твердыми взрывными),
5) идентичная фонемная принадлежность анализируемого гласного (фонема <о> в терминах Московской фонологической школы).

Дополнительно было исследовано произношения слов *хэштег*[3] (предположительно относящегося к той же группе, что и *стоп-кран*) и *хэндаут* (предположительно одноударного).

---

[2] См. про слова *поп группа, стоп-кран, шорт-лист, шорт-трек, ток шоу, бойфренд, полвека*, *бой-баба* [Kalenchuk, Kasatkin, Kasatkina 2012: 608, 813, 922, 830, 44, 592].

[3] *Хэштег* (*хештег*) — слово или фраза, которым предшествует символ #, служащие для объединения групп сообщений по теме или типу в социальных сетях.

В ходе эксперимента информанты зачитывали приведенные ниже предложения с тестовыми словами как в слабой фразовой позиции (не в начале и не в конце синтагмы и не под синтагматическим акцентом), так и в сильной (под фразовым акцентом):

(1) *В конце концов мы решили сами остановить поезд и побежали искать стоп-кран.*
(2) *Когда **стоп-кран** сорвали, раздалось громкое шипение.*
(3) *Когда мы стоптали валенки, пришлось отправиться за галошами.*
(4) *Все сообщения в его блоге объединялись в разные темы при помощи хэштегов.*
(5) *Когда **хэштеги** сняли, всё, конечно, сразу перепуталось.*
(6) *Гарик всегда носил на свои доклады хэндауты, чтобы не делать презентаций.*
(7) *Но однажды **хэндаут** стёрся, и доклад пришлось отменить.*

Тестовые слова подчеркнуты, полужирным шрифтом выделены примеры в слабой фразовой позиции.

В первом эксперименте приняли участие 25 человек в возрасте от 17 до 50 лет, все москвичи, носители литературного произношения, во втором — 15 человек из числа 25, принимавших участие в первом эксперименте.

Прочитанный информантами текст был записан и проанализирован при помощи программ *Speech Analyzer* и *Praat*. В тестовых словах по спектрограммам и осциллограммам были установлены границы гласных для определения их длительности. На рисунках 1–3 приведены сциллограммы и динамические спектрограммы слова *стоп-кран* (сильная позиция, диктор С. А.), *стоп-кран* (слабая позиция, диктор С. А.), *стоптали* (слабая позиция, диктор С. А.).

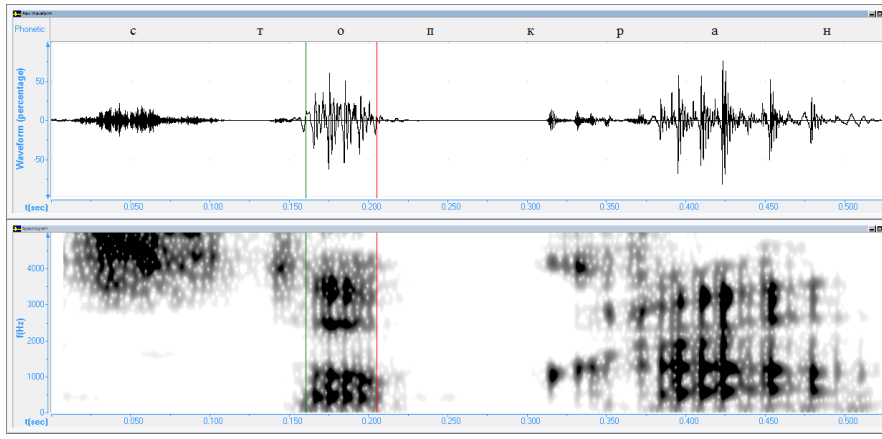


Рис. 1. Осциллограмма и динамическая спектрограмма слова *стоп-кран* (сильная позиция)

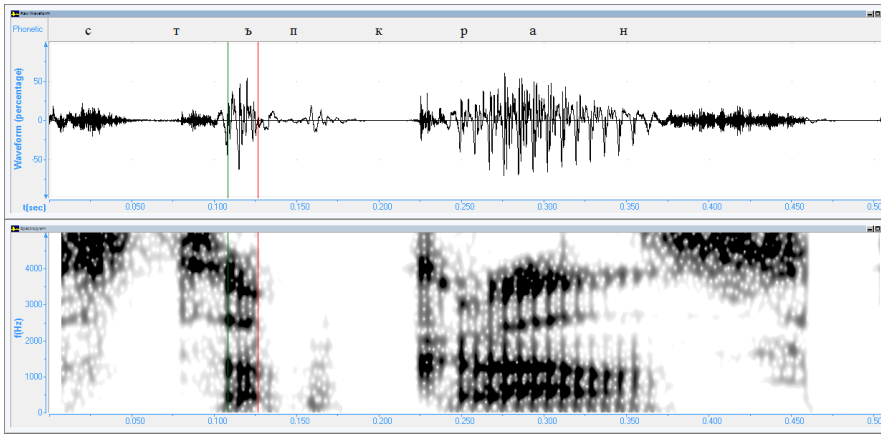


Рис. 2. Осциллограмма и динамическая спектрограмма слова *стоп-кран* (слабая позиция)

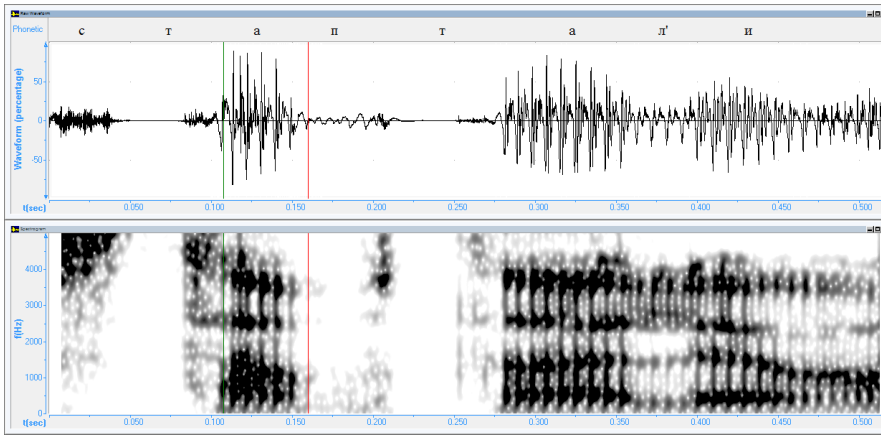


Рис. 3. Осциллограмма и динамическая спектрограмма слова (*с*)*топтали* (слабая позиция)

Данные о длительности гласных в словах *стоп-кран* (в сильной и слабой фразовых позициях) и *стоптать* (в слабой фразовой позиции) приведены в Табл. 1.

Таблица 1. Длительность первого гласного (мс) в словах *стоп-кран* (в сильной и слабой фразовых позициях) и *стоптать* (в слабой фразовой позиции)

информант	*стоп-кран* (**сильная** фразовая позиция)	*стоп-кран* (**слабая** фразовая позиция)	*стоптать* (**слабая** фразовая позиция)
1	43,5	31,9	52,8
2	41,3	22,8	41,7
3	47,7	20,8	50,3
4	47,1	22,8	45,7
5	44,9	30,2	52,0
6	42,2	17,0	33,0
7	37,0	28,6	40,0
8	41,8	29,3	53,7
9	61,4	28,9	48,6
10	50,3	34,5	50,1
11	44,1	36,0	51,3
12	43,9	28,3	48,2
13	45,2	26,1	53,0
14	47,8	33,5	63,5
15	38,1	28,7	50,8
16	48,8	31,4	63,9
17	56,5	19,6	58,5
18	51,2	37,4	67,4
19	53,3	33,6	60,9
20	43,8	28,0	40,0
21	50,5	26,1	46,8
22	51,4	32,9	52,8
23	—	—	—
*24**	*47,6*	*45,1*	*50,5*
25	—	—	—
Итого (дикторы 1–22)	**46,9**	**28,56**	**51,1**

Анализ данных, приведенных в Табл. 1, позволяет сформулировать следующие выводы:

1. В произношении подавляющего большинства информантов (22 из 23)[4] первый гласный в слове *стоп-кран* в слабой фразовой позиции заметно короче как того же гласного в сильной фразовой позиции, так и безударного [а][5] в слабой фразовой позиции (в среднем приблизительно на 40 %).
2. В произношении подавляющего большинства информантов (22 из 23) абсолютная длительность первого гласного в слове *стоп-кран* в слабой фразовой позиции составляет от 17 до 33 мс., что соответствует продолжительности редуцированного гласного, а не гласного полного образования.

Таким образом, исходную гипотезу о принципиальном различии фонетической реализации первого гласного в двусложных составных единицах (типа *стоп-кран*) и обычных фонетических словах (типа *стоптать*) в слабой фразовой позиции можно считать подтвержденной. Отсутствие просодического ядра слова и блокировка правила, в соответствии с которым фонема <о> в первом предударном слоге после парных твердых согласных реализуется звуком [а], свидетельствует о том, что единицы типа *стоп-кран* должны рассматриваться как сочетания двух фонетических слов ([сто́п кра́н], а не [стопкра́н]).

Вторая часть эксперимента была посвящена исследованию произношения слов *хэштег* (предположительно относящегося к той же группе, что и *стоп-кран*) и *хэндаут* (предположительно одноударного). Данные о длительности гласных в словах *хэштег* и *хэндаут* в сильной и слабой фразовых позициях приведены в Табл. 2.

Анализ данных, приведенных в Табл. 2, позволяет сформулировать следующие выводы:

1. В произношении большинства информантов (11 из 14)[6] первый гласный в слове *хэштег* в слабой фразовой позиции заметно короче как того же гласного в сильной фразовой позиции (в среднем приблизительно на 45 %), так и безударного [ы][7] в слабой фразовой позиции.
2. В произношении большинства информантов (11 из 14) абсолютная длительность первого гласного в слове *хэштег* в слабой фразовой позиции составляет от 12 до 46 мс (в среднем — 33,5 мс), что может соответствовать продолжительности редуцированного гласного.

---

[4] Единственный информант, в произношении которого указанного явления не наблюдалось (№ 24), отмечен в таблице звездочкой, полученные от него данные выделены курсивом.

[5] Меньшая длительность ударного [о] по сравнению с безударным [а], по-видимому, объясняется большей собственной длительностью широких гласных по сравнению с гласными среднего подъема [Kuznetsov, Ott 1989].

[6] Информанты, в произношении которых указанного явления не наблюдалось (№№ 8, 9, 14), отмечены в таблице звездочкой, полученные от них данные выделены курсивом. При этом в произношении информантов № 8 и № 9 отмечены нестандартные для данной позиции реализации гласного: [а] и [э].

[7] Меньшая длительность безударного [ы] по сравнению с ударным [е] может объясняться меньшей собственной длительностью закрытых гласных по сравнению с гласными среднего подъема [Kuznetsov, Ott 1989].

Таблица 2. Длительность первого гласного (мс, %) в словах *хэштег* и *хэндаут* в сильной и слабой фразовых позициях

инф	*хэштег* (сильная фразовая позиция)	*хэштег* (**слабая** фразовая позиция)	*хэндаут* (сильная фразовая позиция)	*хэндаут* (**слабая** фразовая позиция)
1	63,7	12,1 (19%)	48,1	33,5 (70%)
2	59,2	33,1 (56%)	?	?
3	61,8	37,3 (60%)	41,2	38,9 (94%)
4	42,6	29,4 (69%)	42,7	46,8 (110%)
5	76,0	27,5 (36%)	42,3	32,0 (76%)
6	44,6	21,1 (47%)	35,1	30,1 (86%)
7	85,5	46,1 (54%)	?	?
8*	*48,6*	*52,0 [а] (107%)*	36,8	39,2 (107%)
9*	*73,5*	*65,1 [э] (89%)*	44,7	49,0 (110%)
10	70,2	44,0 (63%)	75,5	74,9 (99%)
11	42,7	21,5 (50%)	60,1	53,9 (90%)
12	67,6	51,4 (76%)	?	?
13	58,1	44,5 (77%)	62,2	58,7 (94%)
14*	*43,0*	*42,7(99%)*	*47,1*	*44,6 (95%)*
15	–	–	–	–
Итого (дикторы 1–7, 10–13)	**61,1**	**33,5 (55%)**	**48,8**	**45,7 (94%)**

Таким образом, есть основания считать, что для слов *хэштег* и *хэндаут* различия в реализации первого гласного в сильной и слабой фразовых позициях выражены по-разному: у большинства информантов (80%) в слове *хэштег* гласный в слабой фразовой позиции значительно (в среднем почти в два раза) короче, чем в позиции под фразовым акцентом, в то время как в слове *хэндаут* в сильной и слабой фразовых позициях продолжительность первого гласного отличается незначительно и существенно превышает длительность гласного в слове *хэштег* в слабой фразовой позиции. Тем самым, имеются основания считать, что единица *хэштег* представляет собой два фонетических слова (как *стоп-кран*), в отличие от единого фонетического слова *хэндаут*. Представляется, что полученные в ходе настоящего исследования данные могут позволить в некоторых случаях принимать решение об орфографической адаптации заимствований. Так, слово *хэндаут*, несомненно, должно быть записано без пробела, в то время как для второго исследованного в ходе настоящего

эксперимента слова следует рекомендовать раздельное или дефисное написание *хэш тег* или *хэш-тег* (*хеш тег* или *хеш-тег*) по аналогии со случаями типа *поп музыка, стоп-кран*.

В заключение необходимо отметить, что в соответствии с данными аудитивного анализа аналогичные рассмотренным выше слова с широким гласным [а] в первом слоге (*вай фай*, *такс фри*, *фан зона*, *прайс лист*, *хайтек*, *танцпол* и т. п.) ведут себя несколько иначе слов с гласными среднего подъема, закономерности реализации открытого гласного в этих условиях нуждаются в отдельном исследовании.

References

1. *Al'mukhamedova Z.M., Kul'sharipova R.E.* (1980), Vowel reduction and word prosody in Russian dialects with *okan'ye* (Experimental phonetic study) [Reduktsiya glasnykh I prosodiya slova v okayushchikh russkikh govorakh (Eksperim-fonet. Issled.)], Kazan'.
2. *Grammatchikova E. V., Knyazev S. V., Lukyanova L. V., Pozharitskaya S. K.* (2013), Rhythmical structure of phonological word as a function of tonal accent's place in some regional variants of Standard Modern Russian [Ritmicheskaya struktura slova i mesto realizacii tonalnogo aktsenta v regionalnykh variantakh sovremennogo russkogo literaturnogo yazyka], Relevant problems of theoretical and applied phonetics. Papers in honour of O. F. Krivnova [Aktual'nye voprosy teoreticheskoy i prikladnoy fonetiki. Sbornik statey k yuvileyu O. F. Krivnovoy], Moscow.
3. *Grammatchikova E. V., Knyazev S. V.* (2014), Russian posttonic vowels in pretonic position within a prosodic phrase [Russkie zaudarnye glasnye v predudarnoy pozicii vnutri sintagmy], Moscow State University Bulletin, Batch 9 Philology [Vestnik MGU Seriya 9 Filologiya] Vol. 5, pp. 122–134.
4. *Kalenchuk M. L., Kasatkin L. L., Kasatkina R. F.* (2012) The large orthoepic dictionary of Russian. Standard pronunciation and stress in the beginnig of the XXI century: the norm and its variants [Bol'shoy orfoepicheskiy slovar' russkogo yazyka. Literaturnoye proiznosheniye i udareniye nachala XXI veka: morma i ee varianty], Moscow.
5. *Knyazev S. V.* (2007), Phonetic realization of final unstressed vowels in Russian as a cue of prosodic integration of phonological words within a prosodic phrase [Realizatsiya konechnykh zaudarnykh glasnykh v russkom yazyke kak pokazately prosodicheskogo obyedineniya foneticheskikh slov vnutri sintagmy], Phonetics today: Papers presented at the V-th International scientific conference, October 8–10, 2007 [Fonetika segodnya: Materialy dokladov i soobshcheniy V mezhdunarodnoy nauchnoy konferencii 8–10 oktyabrya 2007 g.], Moscow, pp. 110–112.
6. *Knyazev S. V., Urbanovich G. V.* (2002), On the rhythmical model of phonological word in some dialects from Arkhangelsk region [O ritmicheskoj modeli slova nekotorykh arkhangelskikh govorakh], Materials and studies in Russian dialectology [Materialy i issledovaniya po russkoy dialektologii], Moscow.

7. *Kuznetsov V. B., Ott V. A.* (1989) Automatic speech synthesis. Algorithms of "letter-sound" transformation and control of speech segments' duration [Avtomatcheskiy sintez rechi. Algoritmy preobrazovaniya "bukva-zvuk" i upravleniye dlitel'nostyu rechevykh segmentov], Tallinn.
8. *Paufoshima R. F.* (1983) Phonetics of word and phrase in northern Russian dialects [Fonetika slova I frazy v severnorusskikh govorakh], Moscow.
9. *Potebnya A. A.* (1866) On the sound peculiarities of Russian dialects, [O zvukovykh osobennostyakh russkikh narechiy], Philological writings [Filologicheskie zapiski].
10. *Vysotskiy* (1973) On the sound structure of Russian dialects [O zvukovoy structure slova v russkikh govorakh], Studies in Russian dialectology [Issledovaniya po russkoy dialektologii], Moscow.

КОММУНИКАТИВНО-ПРОСОДИЧЕСКИЙ ПОДХОД К ВЫЯВЛЕНИЮ ЭЛЕМЕНТАРНЫХ ДИСКУРСИВНЫХ ЕДИНИЦ В УСТНОМ МОНОЛОГИЧЕСКОМ ТЕКСТЕ[1]

Коротаев Н. А. (n_korotaev@hotmail.com)

РГГУ, РАНХиГС, Институт языкознания РАН,
Москва, Россия

Ключевые слова: устная речь, русский язык, корпус, сегментация, коммуникативная просодия

ELEMENTARY DISCOURSE UNITS IN SPOKEN MONOLOGUES: EVIDENCE FROM COMMUNICATIVE PROSODY

Korotaev N. A. (n_korotaev@hotmail.com)

RSUH, RANEPA, Institute of Linguistics (Russian Academy of Sciences), Moscow, Russia

The paper addresses the issue of spoken discourse segmentation. Using the corpus “Stories of presents and skiing”, I explore the concept of Elementary Discourse Unit (EDU) — a chunk of speech flow defined on both prosodic and syntactic grounds. I propose for a new procedure to establish EDUs’ boundaries. Compared to previous studies, a communicative perspective is added. I introduce the notion of communicative prosodic constituent, as well as a typology of those. It is based on three oppositions: (i) topic vs. comment; (ii) completion vs. transitional-continuity; (iii) main line vs. parenthesis. These oppositions are defined independently of one another and provide for a six-fold classification.

Several remarks should be made here. First, comments and (optionally) topics are found not only in statements, but also in other illocutionary types — such as questions, directives, vocatives, and so on. Second, it is sometimes hard to distinguish between comment constituents that express transitional continuity properties and topic constituents. I show that in some cases, this distinction can be made even though the intonation patterns are quite similar. Third, parenthetic constituents may as well have internal topics and comments.

[1] Работа выполнена при поддержке РНФ, грант № 14-18-03819.

Next, EDUs' boundaries are re-defined as a subset of communicative prosodic constituents' boundaries. Comment constituents always imply EDU's boundaries, while a topic constituent needs a syntactical support to do so. Finally, I provide an analysis of communicative structure and EDUs boundaries in an excerpt from the corpus.

Keywords: spoken discourse, Russian language, corpus, discourse segmentation, communicative prosody

1. Сегментация речевого потока. Элементарная дискурсивная единица

Хорошо известно, что устная речь порождается не непрерывным потоком, а некоторыми «минимальными порциями». Выявление и классификация таких элементарных единиц входит в число центральных задач как теоретического анализа устного дискурса, так и более практической работы по выполнению дискурсивной транскрипции. По большому счету, реализация системы дискурсивной транскрипции, т. е. графической записи наиболее существенных аспектов устной речи, невозможна без систематического учета принципов сегментации. При этом разбиение устного текста на обособленные друг от друга единицы может проходить на различных уровнях речепорождения. Приведем пример из корпуса русской монологической речи «Истории о подарках и катании на лыжах» [2].

(1) *Pic-RUS_06-f_Ski-T*

1.	/Жил-\был ·· один /дядечка,
2.	··· ' ·· эээ ·· один раз он /проснулся утром,
3.	··· *наверное был /–выходной*,,,
4.	···· /поел,
5.	·· и решил покататься на \лыжах.

Каждая строка фрагмента (1) произносится в рамках единого интонационного контура. Между строками имеются паузы (в транскрипте они отмечены поднятыми точками), которые выполняют двоякую функцию. С одной стороны, паузы могут быть обусловлены физиологической необходимостью сделать вдох, с другой стороны, они используются говорящим для планирования

[2] Описание корпуса и полные транскрипты всех входящего в него текстов доступны в Интернете по адресу http://spokencorpora.ru/. Здесь и далее в заголовках примеров указываются кодовые номера текстов. Нумерованные строки соответствуют элементарным дискурсивным единицам. При помощи слешей и стрелок обозначаются движения тона соответственно на ударных и внеударных слогах словоформ-акцентоносителей. Пунктуационные знаки в конце строк передают фазово-иллокутивные значения ЭДЕ. Подробнее см. [Кибрик, Подлесская (ред.) 2009].

дальнейших отрезков дискурса. (В терминах У. Чейфа [Chafe 1994], каждая строка вербализует отдельный «фокус сознания», а паузы необходимы для «перемещения» сознания от одного фокуса к другому.) С точки зрения семантики, в строке заключено описание одного события или состояния; в синтаксической структуре этому каноническим образом соответствует формат простой клаузы. Кроме того, строки входят в определенные смысловые отношения со своими «соседями», формируя таким образом локальную дискурсивную структуру текста. Так, если использовать аппарат теории риторической структуры ([Mann, Thompson 1988]), то между строками 2, 4 и 5 можно усмотреть риторическое отношение Последовательности (Sequence); строка 3 связана со строкой 2 отношением Фона (Background); строка 1 выполняет по отношению ко всему последующему контексту роль Экспозиции (Setting). Подробнее о применении TPC в анализе устных рассказов см. [Литвиненко и др. 2009].

Таким образом, в примере (1) представлен случай координации физиологического, просодического, семантико-синтаксического, когнитивного и риторического аспектов речепорождения. Однако эти факторы, столь различные по своей природе, далеко не всегда действуют согласованно. В частности, в ряде работ (см., среди прочих, [Croft 1995], [Кибрик, Подлесская (ред.) 2009]) было показано, что доля совпадений между единицами, выделяемыми на интонационных основаниях, и простыми клаузами в реальных текстах не превосходит 60–70 процентов. Отсюда следует, что при сегментации устного текста перед транскрайбером стоит выбор: либо производить разбиение на базе какого-то одного критерия, либо пытаться учесть многофакторность анализируемого явления.

Первый подход связан, в первую очередь, с применением понятия *интонационной единицы*. Интонационная единица характеризуется набором формальных просодических признаков — таких как единый контур частоты основного тона, наличие акцентного центра, закономерности в изменении громкости и скорости произнесения, а также наличие пограничных пауз. Взгляд на интонационную единицу как на минимальный шаг в порождении дискурса принят в большом количестве работ, посвященных дискурсивной транскрипции — см., в частности, [Chafe 1994, Du Bois 2000, Izre'el, Mettouchi 2015]. Вопрос о содержательном наполнении выделяемых единиц (в том числе, об их способности вступать в риторические отношения с другими единицами текста) при таком методе отодвигается на второй план.

Второй подход был предложен в коллективной монографии [Кибрик, Подлесская (ред.) 2009] и продолжает развиваться в рамках проекта по изучению устной речи, практические результаты которого представлены на сайте http://spokencorpora.ru/. Центральное понятие этого подхода — *элементарная дискурсивная единица* (далее — ЭДЕ). В отличие от интонационной единицы, ЭДЕ понимается как комплексный объект, обладающий не только просодическими, но и семантико-синтаксическими, и — опосредованно — риторическими свойствами. Алгоритм выявления ЭДЕ в потоке речи, в неявном виде изложенный в [Кибрик, Подлесская (ред.) 2009] и реализованный в дискурсивной транскрипции корпуса «Рассказы о сновидениях», можно сформулировать следующим образом:

(i) производится сегментация текста на простые клаузы;
(ii) фрагменты, которые не удается интерпретировать как части каких-либо простых клауз («лишние куски»), признаются неканоническими ЭДЕ;
(iii) выделенные клаузы тестируются на «просодическую атомарность»: если внутри клаузы проходит ощутимая просодическая граница (т. е., по сути, если в одной клаузе содержится более одной интонационной единицы), она разделяется на две или более ЭДЕ, в противном случае — признается канонической ЭДЕ.

Можно заметить, что этот алгоритм не только позволяет выполнить сегментацию текста, но и задает определенную типологию ЭДЕ. С одной стороны, имеется «золотой стандарт», при котором простая клауза совпадает с интонационной единицей. С другой стороны, возможны различные отклонения от этого канона, связанные либо с дополнительной интонационной фрагментацией клауз, либо с появлением единиц, не вписывающихся ни в одну клаузу. Вместе с тем применение описанного алгоритма сопряжено и с рядом проблем, а для его реализации приходится формулировать не один десяток дополнительных правил — см., в частности, описание подобного рода частных правил для синтаксически сложных конструкций в [Коротаев и др. 2009]. Как представляется, это связано в первую очередь с двумя факторами. Во-первых, чрезмерное значение придается синтаксическому критерию. Во-вторых, интонационный критерий фактически отделен от семантической интерпретации тех последовательностей словоформ, к которым он применяется.

В настоящей работе предлагается альтернативный алгоритм выявления ЭДЕ в потоке речи, разработанный на материале уже упомянутого корпуса «Истории о подарках и катании на лыжах». Синтаксический критерий в нем занимает менее важное место, а на первый план выходит интонационный критерий, «обогащенный» коммуникативной интерпретацией. Ключевую роль в предлагаемом алгоритме играет понятие коммуникативно-просодической составляющей, пояснению которого посвящен раздел 2. Непосредственное описание правил сегментации на ЭДЕ содержится в разделе 3, в котором также представлен иллюстративный фрагмент одного из текстов. В разделе 4 будут подведены краткие итоги проведенного исследования.

2. Коммуникативно-просодические составляющие

Под *коммуникативно-просодической составляющей* мы понимаем цепочку словоформ, выражающую определенное коммуникативное значение и снабженную одним или несколькими коммуникативно релевантными акцентами. В приведенном определении много общего с рассматриваемым в работах Т. Е. Янко понятием коммуникативной составляющей — цепочкой словоформ, служащей носителем одного или нескольких коммуникативных значений, которые, в свою очередь, задаются в том числе и коммуникативно релевантными акцентами [Янко 2001: 19–136]. Важное отличие тут состоит в том, что

коммуникативно-просодическая составляющая не может быть лишена фразового или логического ударения. Соответственно, из числа коммуникативно-просодических составляющих исключаются безударные коммуникативные составляющие, т. е. словоформы или группы словоформ, контекстуально способные выполнять ту или иную коммуникативную функцию (чаще всего — тематическую), но произносимые без акцентов. В подобных случаях мы считаем, что коммуникативная функция не получает просодической поддержки и, в некотором смысле, оказывается нереализованной в дискурсе. При этом входящие в такую безударную группу словоформы включаются в ближайшую коммуникативно-просодическую составляющую.

Поясним сказанное выше на примере.

(2) *Pic-RUS_01-f_Ski-T*

12.	··· [/Человек[3]] [изрядно \↑нажр<u>а</u>лся[4]],
13.	····· ээ [после /\чего—о[3]] ·· [почему-то решил \ещ<u>ё</u>[1] раз покататься на \ л<u>ы</u>жах[1]].
14.	····· [/Покатался[3] он] ·· [не \<u>о</u>чень[1] \↑уд<u>а</u>чно[4]],
15.	·· [так как был \пь<u>я</u>ный[1]],
16.	·· [а за \рулём как известно /\нельз<u>я</u>[2] пить],
17.	··· [/и /попал[5(3)] в \реаним<u>а</u>цию[5(2)]]!

Границы коммуникативно-просодических составляющих размечены при помощи квадратных скобок. Как видно, в элементарной дискурсивной единице может содержаться более одной коммуникативно-просодической составляющей (напомним, что алгоритм членения на ЭДЕ описан ниже, в разделе 3). Для коммуникативно релевантных акцентов посредством верхних индексов указана их принадлежность к типам интонационных конструкций (ИК). Так, например, первая коммуникативно-просодическая составляющая отрывка содержит одну словоформу *человек*, произносимую с восходящим акцентом типа ИК-3; следующая составляющая включает в себя глагольную группу *изрядно нажрался*, акцентоносителем в которой является глагольная форма, произносимая со сложным нисходяще-восходящим акцентом типа ИК-4.

Чаще всего в коммуникативно-просодической составляющей имеется один коммуникативно релевантный акцент, хотя возможны и исключения — см., например, составляющую в строке 17, в которой реализована сложная акцентная фигура ИК-5, подразумевающая сочетание восходящего акцента типа ИК-3 и нисходящего акцента (в данном случае — типа ИК-2). Кроме того, некоторые акценты лишены коммуникативного наполнения: таковы акценты на предложной группе *за рулём* в ЭДЕ 16 и на союзе *и* в ЭДЕ 17. Наконец, в примере (2) представлен и случай безударной коммуникативной составляющей. Речь идет о местоимении *он* в строке 14, которое можно интерпретировать как атоническую тему, расположенную между двумя ударными компонентами коммуникативной структуры — глагольной темой *покатался* и ремой *не очень удачно* (о понятии атонической темы см. [Янко 2001: 73–80]; ниже мы еще

вернемся к нему при обсуждении парентезы). Согласно приведенному выше определению, подобные безударные элементы включаются в одну коммуникативно-просодическую составляющую с той ударной группой, к которой они интонационно примыкают; в данном случае *он* входит в ту же коммуникативно-просодическую составляющую, что и *покатался*.

Наша рабочая гипотеза состоит в том, что, за вычетом случаев самоисправлений и прочих оборванных контекстов, устный текст можно разбить на коммуникативно-просодические составляющие без остатка. Инвентарь основных значений, выражаемых в коммуникативно-просодических составляющих, задается посредством трех противопоставлений:

(i) тема vs. рема;
(ii) завершенность vs. незавершенность;
(iii) основная линия изложения vs. парентеза.

2.1. Тема vs. рема

В контексте иллокуции сообщения (доминирующей в текстах исследуемого дискурсивного жанра) центральное противопоставление коммуникативной семантики понимается нами в «традиционном» ключе. Рема — это конституирующий, обязательный компонент сообщения, закрепляющий за высказыванием принадлежность к данному типу речевого акта; в нем заключено основное содержание сообщения. В свою очередь тема — это неконституирующий, необязательный компонент сообщения, своего рода «зачин». В качестве темы могут выступать именные группы (см. *человек* в строке 12 примера (2)), глагольные единицы (*покатался он* в строке 14), маркеры риторических отношений между фрагментами дискурса (*после чего* в стоке 13) и проч. Стандартное формальное выражение тематических составляющих — восходящие акценты типа ИК-3 или ИК-6.

Формальным выражением ремы часто становится нисходящий акцент типа ИК-1 или ИК-2 — см. строки 13 и 17 примера (2) выше. Впрочем, как мы постараемся показать ниже, в контексте незавершенности рема может акцентироваться и иначе. С синтаксической точки зрения, ремы могут представлять как полные клаузы (обычно это происходит при отсутствии темы), так и части клауз. Кроме того, регулярны контексты, в которых одна клауза разбивается на несколько рематических составляющих — см. пример (3), в котором говорящий постфактум добавляет к глагольной реме *решил покататься на лыжах* еще две непредикативные ремы:

(3) *Pic-RUS_03-m_Ski-T*

1.	[/Один^{6} →чувак] ·· [/решил покататься на \лы̲жах^{1}].
2.	·· [\Ка̲к-то^{1}].
3.	·· [/→Ранним /→зимним \у̲тром^{1}].

Впрочем, нам кажется плодотворной более широкая трактовка понятий «тема» и «рема», согласно которой эти значения реализуются и в других типах речевых актов, допускающих внутреннее коммуникативное членение: вопросах, директивах и проч. (При этом, разумеется, интонационные модели, реализуемые в составляющих таких высказываний, могут существенно отличаться от описанных для сообщения.) В этом отношении мы следуем идеям С. В. Кодзасова, в работах которого номенклатура коммуникативных категорий задается вне привязки к конкретной иллокутивной функции высказывания. См., в частности, определение ремы как «компонента смысла, над которым производится операция при осуществлении речевого акта» в [Кодзасов 1996/2009: 81]. Более того, мы полагаем, что подобное определение применимо и к коммуникативно-просодическим составляющим, формирующим заведомо нерасчлененные речевые акты: например, обращения, акты поиска подходящей вербализации (ударное *ну*), акты «подытоживания» (ударное *вот*) и др.

При формулировании правил сегментации речевого потока на ЭДЕ в разделе 3 мы используем термин «рема» именно в таком, максимально расширенном, понимании.

2.2. Завершенность vs. незавершенность

Если противопоставление темы и ремы непосредственно связано с внутренней коммуникативной организацией высказывания, то значения завершенности / незавершенности отвечают за то, каким образом компоненты высказывания встраиваются в более широкий дискурсивный контекст — см. разграничение между интонацией предложения и интонацией текста в [Янко 2008]. В отличие от темы и ремы, завершенность и незавершенность не задают отрезков в линейной последовательности текста, а также редко имеют отдельное интонационное выражение. Чаще всего эти значения накладываются на тема-рематическое членение высказывания. При этом тематическая составляющая, судя по всему, способна сочетаться только со значением незавершенности, тогда как рема может характеризоваться как завершенностью, так и незавершенностью.

Яркий пример сочетания ремы с незавершенностью представлен в строке 14 примера (2) выше. В составляющей *не \о̲чень*[1] *\↑уда̲чно*[4] реализована аналитическая стратегия кодирования ремы и незавершенности: каждое из этих значений имеет отдельный интонационный коррелят. Нисходящий акцент на *очень* маркирует рему (в данном случае действуют правила выбора акцентоносителя в контексте контраста), а нисходяще-восходящий акцент на *удачно* указывает на незавершенность. Отметим, что если бы говорящий не хотел подчеркнуть незавершенный статус высказывания в рамках всего текста, ему было бы достаточно произнести наречие *удачно* безакцентно: */Покатался*[3] *он не \о̲чень*[1] *удачно*. Подробнее о данной стратегии выражения незавершенности см. [Янко 2008: 131–141, 148–154].

Однако, как уже было отмечено, значительно чаще значения завершенности / незавершенности выражаются совместно со значениями темы / ремы — на тех же словах-акцентоносителях. См., например, строку 12 примера (2), в ко-

торой составляющая *изрядно* \↑*нажр*а*лся*[4] снабжена одним акцентом типа ИК-4. Выбор акцентоносителя производится тут по правилам для рематической составляющей, а движение тона указывает на незавершенность (ср. с потенциальным *изрядно* *нажр*а*лся*[1] в контексте завершенности). При этом, поскольку для незавершенности характерны в целом те же интонационные паттерны, что и для темы, возникает трудноразрешимая омонимия между темой и «ремой с незавершенностью». Так, в рассмотренном только что случае в принципе допустима и тематическая интерпретация, хотя она и представляется нам семантически менее мотивированной: последовательность *изрядно нажрался*, как кажется, является не «зачином» для составляющей *после чего решил ещё раз покататься на лыжах*, а отдельным полноценным сообщением, не в меньшей степени продвигающим дискурс вперед, чем следующая за ним «рема с завершенностью».

Приведем еще один пример, в котором незавершенность выражается совместно с рематическим значением, интонационно «затирая» его.

(4) *Pic-RUS_02-f_Pr-T*

25.	[но –потом когда он узнал /це<u>н</u>у–у[3]],
26.	·· [/он–н] ·· [в \испуге–е] [сказал «/Нет-/нет-\не<u>т</u>[2]],
27.	[это мне не /ну<u>ж</u>но[3]].»,
28.	··· [/и–и мм \↑ушёл],

В строках 26–27 представлен случай полупрямого цитирования. С сегментной точки зрения передача слов персонажа устроена как прямая речь: сохранены все дейктические элементы, использовано экспрессивное утроение *нет*. Составляющая *это мне не нужно* очевидно имеет рематическую функцию: в ней описывается, что именно сообщил персонаж своему собеседнику. Акцентоноситель в этой составляющей также выбран по стандартным для рем правилам. Однако, сегментно исполнив цитацию как прямую, рассказчица интонационно оформила ее как косвенную: об этом говорит восходящее движение тона на *нужно*, придающее всей цитационной конструкции значение незавершенности. Попутно заметим, что в (4) также представлено два случая так называемого нефинального падения. В строках 26 (последняя составляющая) и 28 налицо нисходящее движение тона, характерное для рем, которое, тем не менее, не достигает типичного для данного говорящего уровня частоты основного тона. На наш взгляд, это еще один случай совместного выражения ремы и незавершенности (подробнее см. [Кибрик 2008]).

Имеются и контексты, в которых при наличии восходящего тона безусловно предпочтительна тематическая интерпретация. В частности, это все неглагольные составляющие, а также препозитивные условные придаточные и другие клаузальные единицы с имманентной тематической функцией. Но в целом приходится признать, что разграничение между темой и ремой с незавершенностью нередко базируется на достаточно зыбких основаниях. Этот вопрос нуждается в дальнейшем содержательном анализе.

2.3. Парентеза

Еще одним значением, накладывающимся на противопоставления темы / ремы и завершенности / незавершенности, является парентеза. Под парентезой мы понимаем временный отход от основной линии изложения, которому сопутствует вставка внутрь некоторой дискурсивной последовательности «инородного» фрагмента, характеризуемого внутренней коммуникативной и интонационной цельностью. В примере (2) выше парентеза реализована в строках 15–16: внутрь последовательности *покатался он не очень удачно … и попал в реанимацию* вставлен фрагмент *так как был пьяный, а за рулём как известно нельзя пить*. Отметим, что в [Янко 2001] парентеза рассматривается как подтип атонической темы. В настоящей работе принят иной взгляд на это явление: мы полагаем, что в парентетических фрагментах могут реализовываться те же коммуникативные значения, что и в основной линии изложения. Так, составляющие *так как был пьяный* и *а за рулём как известно нельзя пить* мы квалифицируем как парентетические ремы, совмещенные со значением незавершенности (о незавершенности сигнализирует нефинальное падение)[3].

Таким образом, на основе рассмотренных противопоставлений можно выделить по крайней мере 6 классов коммуникативно-просодических составляющих: темы, ремы с завершенностью, ремы с незавершенностью, а также эти же три класса в контексте парентезы.

3. Правила сегментации на ЭДЕ

Обсудив понятие коммуникативно-просодической составляющей, мы можем перейти непосредственно к правилам сегментации речевого потока на элементарные дискурсивные единицы.

А. Границы ЭДЕ — это подмножество границ коммуникативно-просодических составляющих. Иными словами, одна коммуникативно-просодическая составляющая не может быть разнесена на несколько ЭДЕ. Важное следствие из этого правила — запрет на выделение ЭДЕ, лишенных коммуникативно значимого акцента. Это касается, в частности, и вершинных компонентов в сложноподчиненных конструкциях — в том случае, если они произносятся безударно. Например, в отрывке (5) словосочетание *он не знает*, выступающее в роли вершинной части конструкции с косвенным вопросом, не имеет на себе акцента, а потому объединяется в одну коммуникативно-просодическую составляющую (и, соответственно, в одну ЭДЕ) с придаточной частью:

---

[3] Безусловно, контекст парентезы может оказывать существенное влияние на формальное выражение коммуникативных значений. Например, хорошо известно, что вставочная информация часто произносится в ускоренном темпе и в «смазанном» тональном диапазоне. Однако на наш взгляд, из этого не следует, что в парентезе не может быть собственных тем и рем.

(5) *Pic-RUS_10-f_Pr-T*

11.	[Он не знает чего /вы́брать[3]],
12.	[и /решил[3]] [спросить у \дете́й[1]].

Отметим, что в рамках «синтаксического» алгоритма, описанного в разделе 1, последовательности типа *он не знает* признавались отдельными ЭДЕ.

Б. Правые границы всех рематических составляющих являются правыми границами ЭДЕ — вне зависимости от того, характеризуются ли эти составляющие завершенностью, и вне зависимости от их синтаксического статуса. Это правило было многократно реализовано при разбиении на ЭДЕ приведенных выше отрывков — см. строки 12–14 и 15 примера (2); все строки примера (3); строки 26–28 примера (4); строку 12 примера (5). Кроме того, как было отмечено в разделе 2.1, при расширительной трактовке понятия «рематическая составляющая» в число рем попадают также и единицы, формирующие нерасчлененные речевые акты, отличные от сообщения. Именно на этом основании отдельной рематической составляющей, а значит, и отдельной ЭДЕ, признается дискурсивный макрер *вот* в следующем примере:

(6) *Pic-RUS_05-m_Pr-T*

39.	·· [и /приобрёл[3]] [\→игру́шечный[2] автомобиль],
40.	·· [\во́т[1]].

В. Правая граница тематической составляющей признается правой границей ЭДЕ только в том случае, если в ближайшем правом контексте нет непарентетической составляющей или группы таких составляющих, объединение которой с данной тематической составляющей можно квалифицировать как простую клаузу. В противном случае данная тематическая составляющая не задает правой границы ЭДЕ. Так, тематическая именная группа *один чувак* в строке 1 примера (3) не формирует отдельной ЭДЕ, так как ее объединение с последующей рематической составляющей *решил покататься на лыжах* можно квалифицировать как простую клаузу. То же верно и для тематических составляющих в строках 1–3 отрывка (2). В то же время составляющая *но потом когда он узнал цену* из примера (4) (как кажется, ее наиболее естественно интерпретировать как тему, а не как рему с незавершенностью) задает правую границу ЭДЕ, поскольку ее объединение с правым контекстом заведомо превосходит формат простой клаузы. Отметим, что неочевидный коммуникативный статус этой составляющей в данном случае не влияет на проведение границы ЭДЕ: если считать данную клаузу ремой, она также выделяется в полноценную ЭДЕ, но уже на основании правила Б.

В примере (7) представлены еще два случая, в которых правая граница тематической составляющей совпадает с правой границей ЭДЕ.

(7) *Pic-RUS_01-f_Pr-R*

34.	·· [/\Жена—а[3] в—в \общем-то—о],
35.	·· [не \уверена[1]],
36.	[что она была \↑→счастлива[4] по этому поводу],

Во-первых, в данном примере реализуется стратегия вынесенного топика: тематическая составляющая *жена в общем-то* не может быть объединена в одну ЭДЕ с правым контекстом, поскольку в опорной клаузе происходит повторной заполнение субъектной валентности. Во-вторых, в строке 35 имеется глагольная тема *не уверена* (нисходящий тон в этой составляющей обусловлен принципом адаптации к нисходяще-восходящему акценту в последующей реме), объединение которой с правым контекстом приводит к образованию сложной, а не простой клаузы.

Г. Левые границы парентез являются левыми границами ЭДЕ. Внутри парентез действуют правила сегментации, описанные в пунктах А–В. См. строки 15–16 рассмотренного выше примера (2): совместно они образуют парентезу, при этом каждая из них по отдельности выступает в качестве парентетической рематической составляющей, совмещенной со значением незавершенности.

В заключение приведем несколько более подробную иллюстрацию описанного в настоящей работе подхода к выделению элементарных дискурсивных единиц. Ниже представлен фрагмент транскрипта, в котором дополнительно отмечены границы коммуникативно-просодических составляющих, а в столбце справа содержится краткое описание основных свойств каждой составляющей.

(8) *Pic-RUS_03-m_Pr-T*

16.	···· [Он пошёл в /–магазин[6]],	Рема + незавершенность; правило Б.
17.	···· [присматриваться к \ машинам[1]].	Рема + завешенность (фрагментация инфинитивной конструкции); правило Б.
18.	···· [Когда ему—у одна машина / понравилась[3]],	Клаузальная тема; правило В (есть правая граница ЭДЕ).
19.	···· $_{i}$[/\о—он—н] $_{ii}$[/спросил] $_{iii}$[у /покупателя-я || \о—ой /\ продавца—а[6]],	i. Неглагольная тема; правило В (нет правой границы ЭДЕ). ii. Глагольная тема; правило В (нет правой границы ЭДЕ). iii. Глагольная тема; правило В (есть правая граница ЭДЕ).
20.	[сколько—о ··· она \стоит[2]].	Рема + завершенность; правило Б.
21.	···· $_{i}$[/Он[3]] $_{ii}$[ему—у /сказал[6]],	i. Неглагольная тема; правило В (нет правой границы ЭДЕ). ii. Глагольная тема; правило В (есть правая граница ЭДЕ).

22.	[что она стоит /**много**^{2}].	Рема + завершенность; правило Б.
23.	···· $_{i}$[И—и он решил что—о' их б= || семейный /бюджет^{3}] $_{ii}$[этого не \потянет^{2}].	i. Тема; правило В (нет правой границы ЭДЕ). Группа *и он решил*, которая выделалась бы в отдельную ЭДЕ при использовании «синтаксического» метода, не формирует отдельной коммуникативно-просодической составляющей, а потому не задает правой границы ЭДЕ (правило А). ii. Рема; правило Б.
24.	·· $_{i}$[И в конце /концов—в^{3}] $_{ii}$['он / купил—л^{3}] ·· $_{iii}$[/коллекционную \ машинку^{1}].	i. Неглагольная тема; правило В (нет правой границы ЭДЕ). ii. Глагольная тема; правило В (нет правой границы ЭДЕ). iii. Рема + завершенность; правило Б.
26.	·· $_{i}$[↓–Жене^{1}].	Рема + завершенность (фрагментация именной группы); правило Б.
27.	·· [–Вот^{1}].	Рема в нерасчлененном высказывании + завершенность; правило Б.
28.	·· [–Всё-е^{1}].	Рема в нерасчлененном высказывании + завершенность; правило Б.

4. Выводы

Предлагаемые в настоящей работе правила сегментации речевого потока на элементарные дискурсивные единицы основаны на совместном учете двух факторов: коммуникативно-просодического и синтаксического. Введено понятие коммуникативно-просодической составляющей — цепочки словоформ, выражающей определенное коммуникативное значение и снабженной хотя бы одним коммуникативно релевантным акцентоми. Выделяется шесть классов такого рода составляющих: тема, рема с завершенностью, рема с незавершенностью и эти же классы в контексте парентезы. В алгоритме разбиения устного текста на ЭДЕ зафиксирована неравнозначность коммуникативнно-просодических составляющих при построении локальной дискурсивной структуры. Рематические составляющие вносят значительно более весомый вклад в «продвижение дискурса вперед», чем тематические. Поэтому правая граница рематической составляющей всегда соответствует правым границам ЭДЕ; тогда как тема задает правую границу ЭДЕ только в том случае, если имеется необходимая синтаксическая «поддержка».

Литература

1. *Кибрик А. А.* (2008) Есть ли предложение в устной речи? // Фонетика и нефонетика. К 70-летию Сандро. В. Кодзасова. М.: ЯСК, 104–115.
2. *Кибрик А. А., Подлесская В. И.* (ред.) (2009) Рассказы о сновидениях: Корпусное исследование устного русского дискурса. М.: ЯСК.
3. *Кодзасов С. В.* (1996) Законы фразовой акцентуации // Просодический строй русской речи. М., 181–204. Перепечатано в [Кодзасов 2009].
4. *Кодзасов С. В.* (2009) Исследования в области русской просодии. М. ЯСК.
5. *Коротаев Н. А., Кибрик А. А., Подлесская В. И.* (2009) Осложнения канонической структуры: на стыке моно- и полипредикативности // Кибрик А. А., Подлесская В. И. (ред.) Рассказы о сновидениях: Корпусное исследование устного русского дискурса. М.: ЯСК, 219–332.
6. *Литвиненко А. О., Подлесская В. И., Кибрик А. А.* (2009) Анализ рассказов о сновидениях с точки зрения иерархической структуры дискурса // Кибрик А. А., Подлесская В. И. (ред.) Рассказы о сновидениях: Корпусное исследование устного русского дискурса. М.: ЯСК, 431–463.
7. *Янко Т. Е.* (2001) Коммуникативные стратегии русской речи. М.: ЯСК.
8. *Янко Т. Е.* (2008) Интонационные стратегии русской речи в сопоставительном аспекте. М.: ЯСК.
9. *Chafe, W.* (1994) Discourse, consciousness, and time. Chicago: University of Chicago Press.
10. *Croft, W.* (1995) Intonation units and grammatical structure // Linguistics 33, 839–882.
11. *Du Bois, J.* (dir.) (2000) Santa Barbara Corpus of Spoken American English. Pt 1–4. Philadelphia: Linguistic Data Consortium.
12. *Izre'el, S., Mettouchi, A.* (2015) Representation of speech in CorpAfroAs: Transcriptional strategies and prosodic units // Mettouchi, Amina, Martine Vanhove and Dominique Caubet (eds.), Corpus-based Studies of Lesser-described Languages: The CorpAfroAs corpus of spoken AfroAsiatic languages. Amsterdam: John Benjamins, 13–41.
13. *Mann, W. C., Thompson, S. A.* (1988) Rhetorical structure theory: toward a functional theory of text organization // Text, 8, 243–281.

References

1. *Chafe W.* (1994), Discourse, consciousness, and time, University of Chicago Press, Chicago.
2. *Croft W.* (1995), Intonation units and grammatical structure, Linguistics 33, pp. 839–882.
3. *Du Bois J.* (dir.) (2000), Santa Barbara Corpus of Spoken American English, Pt 1–4, Linguistic Data Consortium, Philadelphia.

4. *Izre'el S., Mettouchi A.* (2015) Representation of speech in CorpAfroAs: Transcriptional strategies and prosodic units, in Mettouchi, Amina, Martine Vanhove and Dominique Caubet (eds.), Corpus-based Studies of Lesser-described Languages: The CorpAfroAs corpus of spoken AfroAsiatic languages, John Benjamins, Amsterdam, 13–41.
5. *Kibrik A. A.* (2008), Is sentence viable in spoken discourse? [Est' li predlozhenie v ustnoy rechi?], Phonetics and non-phonetics. A 70th birthday Festshcrift for Sandro V. Kodzasov [Fonetika i nefonetika. K 70-letiyu Sandro V. Kodzasova], Languages of Slavonic Culture, Moscow, pp. 104–115.
6. *Kibrik A. A., Podlesskaja V. I.* (eds.) (2009), Night Dream Stories: A corpus study of spoken Russian discourse [Rasskazy o snovideniyakh: Korpusnoe issledovanie ustnogo russkogo diskursa], Languages of Slavonic Culture, Moscow.
7. *Kodzasoc S. V.* (1996), The laws of phrasal accentuation [Zakony frozovoy aktsentuatsii], in Prosodic structure of Russian speech [Prosodicheskiy story russkoy rechi]. Moscow. Reprinted in Kodzasov 2009.
8. *Kodzasov S. V.* (2009), Studies in Russian Prosody [Issledovaniya v oblasti russkoy prosodii], Languages of Slavonic Culture, Moscow.
9. *Korotaev N. A., Kibrik A. A., Podlesskaja V. I.* (2009), Complex structures: between simple predication and polypredication [Oslozhneniya kanonicheskoy struktury: na styke mono- i polipredikativnosti], in Kibrik A. A., Podlesskaja V. I. (eds.), Night Dream Stories: A corpus study of spoken Russian discourse [Rasskazy o snovideniyakh: Korpusnoe issledovanie ustnogo russkogo diskursa], Languages of Slavonic Culture, Moscow, pp. 219–332.
10. *Litvinenko A. O., Kibrik A. A., Podlesskaja V. I.* (2009), Night dream stories from a hierarchical structure perspective [Analiz rasskazov o snovideniyakh s tochki zreniya ierarkhicheskoy struktury diskursa], in Kibrik A. A., Podlesskaja V. I. (eds.), Night Dream Stories: A corpus study of spoken Russian discourse [Rasskazy o snovideniyakh: Korpusnoe issledovanie ustnogo russkogo diskursa], Languages of Slavonic Culture, Moscow, pp. 431–463.
11. *Mann W. C., Thompson S. A.* (1988), Rhetorical structure theory: toward a functional theory of text organization, Text, 8, pp. 243–281.
12. *Yanko T. E.* (2001), Communicative strategies of Russian speech [Kommunikativnye strategii russkoy rechi], Languages of Slavonic Culture, Moscow.
13. *Yanko T. E.* (2001), Intonation strategies of Russian speech from a contrastive perspective [Intonatsionnye strategii russkoy rechi v sopostavitel'nom aspekte], Languages of Slavonic Culture, Moscow.

ФУНКЦИОНАЛЬНЫЙ АНАЛИЗ НЕВЕРБАЛЬНОГО КОММУНИКАТИВНОГО ПОВЕДЕНИЯ[1]

Котов А. А. (kotov@harpia.ru),
Зинина А. А. (zinina_aa@nrcki.ru)

НИЦ «Курчатовский институт», Москва, Россия

Ключевые слова: коммуникативные стимулы, невербальная коммуникация, мультимодальный корпус, эмоциональные компьютерные агенты

FUNCTIONAL ANALYSIS OF NON-VERBAL COMMUNICATIVE BEHAVIOR

Kotov A. A. (kotov@harpia.ru),
Zinina A. A. (zinina_aa@nrcki.ru)

National Research Center “Kurchatov Institute”, Moscow, Russia

In this study we represent functional annotation of the Russian Emotional Corpus (REC). The annotation is appended to the regular annotation of eyes, eyebrows and hand movements with supplementary annotation for head and corpus movements. The annotation records communicative functions, where a movement is intended for a particular goal or can be understood as connected to a particular goal/stimulus by the addressee. We show that a particular function can be expressed by different patterns, utilizing facial expression and/or hand/body movements. Functional annotation is also used as a non-terminal symbol in a generative grammar to produce non-verbal behavioral patterns.

Keywords: communicative stimuli, nonverbal communication, multimodal corpora, emotional computer agents

[1] Исследование выполнено при поддержке гранта РФФИ 15-29-01173.

1. Введение

С точки зрения лингвистики, невербальное коммуникативное поведение (НКП) обладает интересной спецификой: с одной стороны, оно не является основным каналом коммуникации и в малой степени контролируется говорящим (по сравнению с речевым потоком), поэтому содержит меньше произвольных языковых знаков. С другой стороны, НКП необходимо для поддержания личной коммуникации: элементы НКП могут передавать информацию об эмоциональном состоянии адресанта, об отношении к предмету сообщения, передавать информацию, связанную с прагматикой сообщения.

При анимировании трёхмерных компьютерных персонажей или роботов, общающихся с человеком, использование НКП становится необходимым условием создания привлекательного персонажа. Мультимодальные корпуса при этом служат основной исследовательской базой: на основе корпусов разрабатывается модель коммуникативного поведения для эмоциональных агентов, корректность которой исследуется в серии последующих экспериментов (Rehm, André, 2008). Синтез поведения, таким образом, является методом верификации избранной для корпуса схемы разметки. Необходимость ориентироваться на синтез НКП также требует описывать в разметке корпуса «глубинные» стимулы, которые в дальнейшем позволят синтезировать наблюдаемое поведение.

Многие проекты корпусов используют для такой глубинной разметки названия эмоций, прежде всего, инвентарь «базовых эмоций» П. Экмана: удивление, страх, отвращение, гнев, радость и печаль (Ekman, Friesen, 1975). Например, некоторое расширение этого набора используется при создании корпуса GEMEP, где актёры разыгрывают каждую из эмоций предложенного инвентаря (Bänziger, Scherer, 2007). Хотя «разыгранные» корпуса предоставляют достаточно чистые портреты каждой эмоции, они, вместе с тем, не вполне применимы к описанию и синтезу коммуникативного поведения, где смешиваются выражение эмоций и коммуникативных намерений. Для этих целей требуется исследование материала эмоциональных диалогов, например, телевизионных ток-шоу в корпусе EmoTV (Martin, Devillers, 2009) или кинофильмов в Мультимедийном русском корпусе (Гришина, 2014). Анализ такого материала позволяет анализировать ряд важных процессов, характерных для естественной коммуникации: во-первых, отмечать смешанные эмоции (blended emotions), к примеру, одновременное проявление в поведении человека гнева и отчаяния (Martin, Devillers, 2009), а во-вторых, связывать коммуникативные действия не только с эмоциями, но и с коммуникативными целями собеседников. Так, в работе (Гришина, 2014) жесты соотнесены с поверхностными или с глубинными иллокутивными силами высказывания: это позволяет описывать конфигурации жестов как в зависимости от семантики высказывания (например, перебирание пальцами для обозначения перемещения), так и в зависимости от коммуникативных целей (например, жест «дуги» для маркирования вопроса).

В данной работе мы представляем новый блок разметки Русского эмоционального корпуса (Котов, 2009, 2014), фиксирующий коммуникативные функции элементов НКП для задачи последующего синтеза коммуникативного поведения.

2. Функциональная разметка

В корпусе REC собраны видеозаписи реальных эмоциональных диалогов (устных университетских экзаменов и диалогов в муниципальной службе одного окна). Материал сопровождается разметкой — записываются: (а) речь участников коммуникации с размеченной прагматической информацией, (б) мимика и относительно длительные (более 0,5 с) изменения направления взгляда, (в) жесты и иные движения, выполняемые руками.

Элементы коммуникативного поведения (или их комбинации — паттерны) могут передавать информацию в коммуникации, в этом случае мы считаем, что такие элементы обладают некоторой коммуникативной функцией. При этом элемент НКП может быть вызван определённым стимулом и непроизвольно передавать информацию в коммуникации, либо может сознательно использоваться адресантом для достижения коммуникативной цели.

В реальном материале эти два случая часто трудно разделить: демонстрируя удивление (например, расширяя глаза, поднимая брови и открывая рот) говорящий может не только непроизвольно выражать эмоцию, но также пытаться поставить под сомнение высказывание адресата (как бы говоря *Я не понимаю, что за ерунду ты говоришь!* — см. пример 1), оправдывать собственное бездействие (*~Я не понимаю, что ты от меня хочешь!* — см. пример 2) или потерю социального лица при ответе (*~Я не понимаю, как я мог такое сказать!* или *Я не понимаю, что тут ещё можно ответить!* — см. пример 3). Сходный паттерн также используется в качестве эмфазы, если говорящий стремится показать тривиальность ответа, для чего он с помощью «удивления» демонстрирует некоторую неадекватность вопроса (*~Это так просто! Удивительно, что ты меня об этом спрашиваешь!* — см. пример 4). П. Экман в своей работе (Ekman, 2003) показывает, что при проявлении эмоции изумления больше задействован рот, тогда как при удивлении, связанном с вопросом, больше используются брови — таким образом, внешний паттерн оказывается связан с коммуникативной функцией эмоции.

(1) Собеседник: *А вот госпожа* <называет фамилию>. [*Судя по списку*] *даже у меня занималась.*
Информант: *Что?* (смотрит на собеседника, приближается, прищуривается — в данном случае используется другой паттерн удивления: сосредоточенное рассматривание удивительного объекта)[2]
Собеседник: *Я вас даже не помню.*

(2) Собеседник: *Бумажки на стол кладите ещё!*
Информант: *Вот сюда?* (замирает — останавливает жесты, расширяет глаза, переводит взгляд на адресата)[3]

---

[2] Фрагмент 20080717-c15, 00:08.

[3] Фрагмент 20081230-a24-fr, 00:30.

(3) Информант (предлагает вариант ответа): *Это не это — «глух-глуп»?*
Собеседник: *Это не антономазия.*
Информант: *Да что ж такое?* (всплескивает руками, поправляет волосы, далее — поднимает брови и открывает рот)[4]

(4) Собеседник: *А грамматически, грамматически здесь что меняется по сравнению с…?*
Информант: *Время! Время, естественно!* (поднимает плечи и брови на фразовых ударениях, расширяет глаза, переводит взгляд на собеседника)[5]

В приведённых случаях мы имеем дело с «адресованными эмоциями» или pull-эмоциями: их внешнее выражение может быть усилено говорящим, чтобы служить достижению особых коммуникативных целей: оказать воздействие на адресата или снизить ущерб собственному социальному лицу от допущенных ошибок. Как видно из примеров, сходные паттерны выражения адресованных эмоций могут использоваться в различных коммуникативных функциях, при этом для синтеза коммуникативного поведения необходимо решить обратную задачу — для некоторой коммуникативной функции получить список возможных адресованных эмоций (или иных паттернов), которые могли бы выразить эту функцию в данном контексте. В результате, если говорящий стремится показать неадекватность входящей просьбы со стороны собеседника, он может демонстрировать «адресованное удивление» (аналогично примеру 2), «адресованное отвращение» (например, морщить нос), адресованную усталость (например, выдувать воздух), дистанцирование в коммуникации, либо «неэмоциональные» коммуникативные знаки — различные знаки остановки собеседника, знаки несогласия или знаки отсутствия возможности.

Функциональная разметка в корпусе REC направлена на решение этой задачи и обеспечивает поиск паттернов НКП, выражающих определённую коммуникативную функцию. Коммуникативные функции размечаются в корпусе REC для всех движений, которые (а) намеренно выполняются адресантом для достижения некоторой коммуникативной цели или (б) непроизвольно выполняются адресантом, но могут быть достаточно однозначно поняты собеседником как симптом некоторой эмоции, намерения или внутреннего состояния адресанта.

Коммуникативные функции выбираются из списка и приписываются мимике и движениям рук из существующей разметки корпуса в тех случаях, где функция является достаточно определённой. Разметка движений головы и тела в корпусе ограничивается функциональной разметкой, иными словами, кивки, подъёмы и повороты головы размечаются только как, например, «понимание», «указание на референт» или «отрицание». Не размечаются движения, обслуживающие изменение направления взгляда или жесты (поворот головы вслед за взглядом, наклон головы, чтобы почесать ухо и т. д.).

---

[4] Фрагмент 20081225-zh-b2, 04:11.

[5] Фрагмент ah_07, 09:59.

Для разметки всех типов движений в корпусе используется общий список значений (тэгов) — см. Таблицу 1.

Таблица 1. Коммуникативные функции в разметке корпуса REC

Коммуникативная функция		**Описание**
Понимание, согласие, одобрение		знаки, выражающие понимание адресантом слов собеседника, согласие с этими словами или одобрение действий собеседника
Отрицание, несогласие, возражение		выражение отрицания, знаки несогласия со словами собеседника, обозначение исправлений к собственной речи
Апелляция		знаки, направленные на усиление воздействия на собеседника некоторого аргумента, высказанного в речи
Побуждение		знак собеседнику начать действовать или принять решение
Ожидание обратной связи		краткий вопрос собеседнику, переспрос, запрос обратной связи
Остановка адресата		знаки, направленные на остановку речи или действий собеседника
Отсутствие, невозможность		демонстрация отсутствия какого-либо объекта или демонстрация невозможности действовать; например, информант разводит руками, сжимает губы
Обратная связь		краткий знак понимания слов собеседника
Привлечение внимания		знаки, направленные на привлечение внимания собеседника к адресанту или к объекту речи
Демонстрация непонимания		знаки, выражающие непонимание адресанта и направленные на воздействие в отношении собеседника: попытки заставить собеседника повторить вопрос, проявить снисхождение и т. д.
Демонстрация неадекватности адресата		знаки, выражающие отрицательную оценку слов или действий собеседника, их неадекватность
Коммуникативные функции, связанные с выражением внутренних состояний	я:воодушевление	знаки, выражающие подготовку адресантом к действию, или демонстрация готовности действовать
	я:расслабление, удовлетворение	проявление реакции удовлетворения или облегчения адресанта в результате некоторого позитивного события
	я:смущение, фрустрация	проявление смущения или фрустрации адресанта, вызванных некоторым внешним событием (если действия вызваны собственным неадекватным поведением адресанта, то используется аннотация «компенсация, закрытие»)

Коммуникативная функция		**Описание**
Коммуникативные функции, связанные с выражением внутренних состояний	я:хезитация	проявление сомнений и колебаний адресанта во всех случаях, не связанных с решением задачи (в этом случае используется «я:размышление») или выбором из нескольких явных альтернатив («выбор варианта»)
	я:радость	проявление активной реакции адресанта на позитивное событие
	я:умиление, сочувствие	выражение адресантом умиления или сочувствия в отношении собеседника или объекта речи
	я:размышление	коммуникативные действия адресанта, связанные с процессом размышления или решения задачи
Коммуникативные функции, связанные с воздействием на адресата	ты:позитивн к объекту	попытка вызвать у собеседника позитивные чувства по отношению к объекту или третьему участнику ситуации
	ты:позитивн к адресанту	попытка вызывать у собеседника симпатию к адресанту, кокетство с собеседником
	ты:сочувствие к адресанту	попытка вызвать у собеседника сочувствие или снисхождение по отношению к адресанту, попытка снизить давление (действие, угрожающее социальному лицу, или FTA) со стороны собеседника
	ты:пренебрежение	попытка снизить важность события или объекта для собеседника, невербальный аналог высказывания *Это всё ерунда!*
	ты:попытка успокоить адресата	попытка снизить остроту негативных переживаний собеседника
	ты:негативн	попытка спровоцировать негативные переживания собеседника, например, намерение устыдить или напугать
Дистанцирование		значимое увеличение дистанции в коммуникации, отклонение адресанта назад с целью демонстрации неприятия, непонимания или неадекватности адресата; данная аннотация используется, если действию трудно приписать более конкретную функцию (отрицание, несогласие, возражение; демонстрация неадекватности адресата; я:смущение; фрустрация)
Указание на объект		мимический паттерн, движение глаз, головы или тела с целью указания на физический объект
Коммуникативные функции, сопровождающие речь	речь:эмфаза	знаки выделения сегментов речи, фразового ударения, дополнительные знаки подтверждения собственных слов, например, кивки после ассертивного высказывания
	речь:референт	различные средства обозначения референтов

Коммуникативная функция		**Описание**
Коммуникативные функции, сопровождающие речь	речь:операция с референтом	знаки, демонстрирующие преобразование или перемещение референта
	речь:замена референта	обозначение перехода в речи к другому референту, смена темы
Выбор варианта		выражение колебаний при выборе из альтернатив; альтернативы должны быть заданы явно, например, два варианта действий или несколько билетов на столе (остальные колебания размечаются как «я:хезитация»)
Стимулирование		действия, подталкивающие адресанта или собеседника к ответу или решению задачи, например, круговые махи ладонью перед грудью, невербальные аналоги высказываний *Ну!* или *Давай-давай!*
Компенсация, закрытие		действия, с помощью которых адресант скрывает глаза, рот, другие части лица; улыбки и смех, направленные на снижение категоричности речевого высказывания
Другое		все остальные случаи, где коммуникативная функция представляется достаточно определённой, но отсутствует в приведённом списке

Используя общий список категорий для разметки различных элементов НКП, мы предполагаем, что одна и та же коммуникативная функция может проявляться с помощью различных исполнительных органов: мимики, жестов, движений головы и тела. Как видно из Таблицы 2, это верно не для всех коммуникативных функций. Существуют определённые тенденции или ограничения на выражение коммуникативных функций. Так, дистанцирование проявляется преимущественно через движения корпуса тела (в 90 % случаев) и иногда через движения головы (10 %), но не через движения рук. Попытка вызвать симпатию или сочувствие к адресанту, выражение смущения, размышления, либо хезитация передают эмоциональную информацию и выражаются преимущественно через мимику (77–88 % случаев). Согласие, отрицание и эмфаза передаются преимущественно движениями головы (69–82 % случаев), а «стимулирование» и «закрытие» выполняются преимущественно руками: человек машет по кругу ладонью, стремясь «подогнать» процесс рассуждения, либо демонстрирует закрывающие жесты. Однако данная таблица также показывает, что даже при наличии тенденций, подавляющее большинство коммуникативных функций распределены между разными исполнительными органами. Так, запрос (ожидание) обратной связи может выполняться (а) с помощью тела и мимики (64 % случаев), например, с помощью приближения к собеседнику, прищуренного взгляда или поднятия бровей, (б) с помощью движений головы (28 % случаев), например, движением подбородка вверх, вбок, изменением наклона головы («перекладыванием» головы с левого на правое плечо) или демонстративным поворотом лица вбок при сохранении прямого взгляда, а также (в) с помощью

жестов: открытой вверх ладони, пальцами направленной к собеседнику, либо с помощью демонстративного замедления или остановки выполняемого жеста.

В Таблице 2 мы также можем видеть результаты изменения (или редукции) общего паттерна, который проявляется то через одни, то через другие исполнительные органы. Например, «отсутствие, невозможность» наряду с поворотами головы (аналогично жесту «*Нет!*») может выражаться комплексным паттерном: человек поднимает плечи, закрывает глаза, поднимает брови и обращает открытые ладони к адресату. В реальном поведении мы можем наблюдать отдельные элементы этого сложного паттерна: человек может только поднимать брови, только пожимать плечами или только демонстрировать ладони. Возможны и варианты паттерна, например, элемент «пожимать плечами», обычно выполняемый без вздоха, может по-разному комбинироваться с дыханием: информант может (а) демонстративно вдыхать, поднимая плечи, (б) вдыхать и выдыхать, как бы показывая объём требуемых усилий, или (в) демонстративно выдыхать, опуская плечи, как бы показывая усталость и невозможность выполнить действие.

Таблица 2. Выражение коммуникативных функций через различные органы тела (для наиболее частотных функций)

	Коммуникативная функция	**Мимика или тело**		**Голова**		**Руки**		**Всего**
		n	**%**	**n**	**%**	**n**	**%**	**n**
Функции, выражаемые преимущественно мимикой и телом	дистанцирование	45	**90,0**	5	10,0	0	0,0	**50**
	ты:позитивн к адресанту	71	**88,8**	9	11,3	0	0,0	**80**
	ты:сочувствие к адресанту	31	**88,6**	4	11,4	0	0,0	**35**
	я:смущение-фрустрация	251	**83,9**	38	12,7	10	3,3	**299**
	я:размышление	330	**77,8**	57	13,4	37	8,7	**424**
	я:хезитация	72	**77,4**	11	11,8	10	10,8	**93**
	демонстрация непонимания	39	**66,1**	17	28,8	3	5,1	**59**
	ожидание обратной связи	229	**64,1**	101	28,3	27	7,6	**357**
	я:воодушевление	32	**58,2**	10	18,2	13	23,6	**55**
	ты:пренебрежение	33	**54,1**	5	8,2	23	37,7	**61**
Функции, выражаемые преимущественно головой	понимание, согласие, одобрение	21	14,1	123	**82,6**	5	3,4	**149**
	речь:эмфаза	212	19,1	773	**69,8**	123	11,1	**1,108**
	отрицание, несогласие, возражение	24	24,0	69	**69,0**	7	7,0	**100**
	речь:референт	5	7,2	42	**60,9**	22	31,9	**69**
	апелляция	46	32,6	68	**48,2**	27	19,1	**141**

	Коммуникативная функция	Мимика или тело		Голова		Руки		Всего
		n	%	n	%	n	%	n
Функции, выражаемые преимущественно руками	стимулирование	4	3,3	16	13,0	103	**83,7**	**123**
	компенсация, закрытие	224	34,3	15	2,3	414	**63,4**	**653**
	остановка адресата	6	10,2	21	35,6	32	**54,2**	**59**
	отсутствие, невозможность	10	34,5	6	20,7	13	**44,8**	**29**
	другое	191	46,7	198	48,4	20	4,9	**409**
	Всего	**1,876**		**1,588**		**889**		**4,353**

3. Проблема генерации поведения

Использование коммуникативных функций может служить важным элементом для генерации коммуникативного поведения в компьютерных архитектурах. Мы предприняли попытку построить цепочки элементов НКП с помощью контекстно-свободной грамматики, аналогично тому, как в генеративной грамматике строятся высказывания естественного языка. Коммуникативная функция в этом случае является нетерминальным символом, который при порождении поведения заменяется одним из паттернов — терминальных символов грамматики. Прототип грамматики для ситуации ответа на вопрос реализован в среде Prolog. В целом, в ситуации вопроса адресант может сразу отказаться отвечать или попытаться найти и сообщить ответ, что записывается следующим порождающим правилом:

```
вопрос --> отказ; цель _ найти _ и _ сообщить _ ответ.
```

«Отказ» — нетерминальный символ, который может быть выражен в терминальных символах коммуникативной функции «отрицание» в речи (адресант говорит *Нет!*) и в жестах (адресант крутит головой или отрицательно машет рукой), а также сопровождаться различными стратегиями отказа — адресант может показать, что ему очень тяжело, что он ничего здесь не понимает, либо предложить ответить совсем на другой вопрос:

```
отказ --> ком _ функ _ отрицание, стратегии _ отказа _ 01.

ком _ функ _ отрицание --> [текст _ нет _ , крутит _
      головой _ 01, машет _ рукой _ 01].

стратегии _ отказа _ 01 --> стратегия _ ты _
      затруднение; стратегия _ ты _ непонимание;
      стратегия _ предложить _ альтернативу.
```

В свою очередь, каждая из стратегий является нетерминальным символом и может порождать элементы из инвентаря коммуникативных функций, которые далее порождают цепочки терминальных элементов поведения — ссылок на движения, наблюдаемые в корпусе и соответствующие данным коммуникативным функциям. В результате, строится множество цепочек возможного коммуникативного поведения; один из примеров приведён в Таблице 3.

Таблица 3. Пример работы порождающей модели коммуникативного поведения

`S = [останавливающий _ жест _ 01, текст _ подождите _ ,`	Готовясь к ответу, останавливает адресата жестом и речью; элементы порождены коммуникативной функцией *остановка адресата*
`облизывается _ 01, побнимает _ брови _ 01,`	Готовится отвечать: *я:воодушевление*
`чешет _ голову _ 01, дистанцируется _ 01, смотрит _ вбок _ 01,`	Размышляет над вопросом: *я:размышление*
`облизывается _ 01,`	Готовится произнести ответ: *я:воодушевление*
`текст _ ответ _ ,`	Произносит ответ
`отворачивается _ 01, закрывает _ глаза _ 01,`	Смущается из-за неуверенности в ответе: *компенсация, закрытие*
`машет _ рукой _ 02, морщит _ нос _ 01,`	Старается преуменьшить значимость своего ответа: *ты:пренебрежение*
`вход _ полож _ обр _ связь, улыбается _ 01, хлопает _ в _ ладоши _ 01]`	Получает положительную обратную связь от собеседника и демонстрирует радость: *я:радость*

Конечно, с точки зрения современного уровня анимации компьютерных персонажей, предложенная модель является весьма упрощённой. Вместе с тем, мы рассматриваем её в качестве иллюстрации порождающего подхода к генерации поведения, где предложенный инвентарь коммуникативных функций является важным уровнем «грамматики» коммуникативного поведения и, кроме того, позволяет верифицировать разметку, используемую в корпусе.

4. Совмещение коммуникативных функций в поведенческих паттернах

Функциональная разметка позволяет искать в корпусе сложные поведенческие паттерны, где различные органы тела преследуют разные коммуникативные функции. Одним из таких случаев является компенсация: говорящий вынужден компенсировать иконический жест или прямой взгляд, направленный на адресата, с помощью улыбок или закрывающих жестов (Котов, 2013).

Другая группа сложных примеров показывает, что говорящий может одновременно останавливать «негативные» действия адресата и пытаться оказать на собеседника «позитивное» воздействие.

(5) Собеседник (исправляет информанта): *Тут не метафора.*
Информант: *Сейчас.*
Манипулирует волосами (руки-функция: *компенсация, закрытие*) и закрывает лицо руками (руки-функция: *компенсация, закрытие*), улыбается, закрывает глаза, смотрит вверх (функция: *я:размышление*).
Информант: *Ну хорошо! Тогда* [*я отвечу*] *про путь! Этот пример я точно помню.*
Направляет открытую ладонь к адресату пальцами вверх (руки-функция: *остановка адресата*), улыбается (функция: *ты:позитивн к адресанту*), сильно отклоняется назад (функция: *дистанцирование*)[6]

В первой части примера информант останавливает адресата в речи (*Сейчас!*), при этом демонстрирует закрывающие жесты и размышление. Во второй части примера информант предлагает в речи заменить задание, при этом жестами останавливая возможные возражения со стороны адресата, провоцируя симпатию адресата (улыбаясь и кокетничая) и одновременно дистанцируясь.

(6) Информант: *Ааа.*
Открывает рот и облизывается, двигает губами (функция: *я:размышление*).
Информант: *Сейчас я.*
Машет указательным пальцем к адресату (руки-функция: *остановка адресата*), машет пальцем (руки-функция: *стимулирование*), сжимает губы, выдувает воздух, смотрит вверх (функция: *демонстрация непонимания*), наклоняет голову вниз (голова-функция: *я:размышление*), трет пальцем переносицу (руки-функция: *я:размышление*)[7]

В речи и жестах информант останавливает адресата, при этом демонстрируя активное размышление.

Как показывают «сложные» примеры, информанты могут одновременно останавливать негативную тенденцию (предотвращая возможную критику адресата), провоцировать позитивную тенденцию (кокетничать, вызывая симпатию адресата), регулировать собственные состояния (стимулировать размышление над вопросом), а также компенсировать в поведении собственные коммуникативные действия: демонстрировать закрывающие жесты и улыбки, снижающие ущерб социальному лицу адресата от основных коммуникативных действий говорящего.

---

[6] Фрагмент 200081219-zhurn-a03, 1:04.

[7] Фрагмент 20081225-zhurn-b3, 5:26.

Моделирование элементов НКП в примерах, комбинирующих различные коммуникативные функции, может позволить создать реалистичную компьютерную модель НКП для виртуальных компьютерных агентов и роботов, взаимодействующих с человеком.

Литература

1. *Гришина Е. А.* Жесты и грамматические характеристики высказывания // Мультимодальная коммуникация: теоретические и эмпирические исследования. — М.: «Буки Веди», 2014. — С. 25–47.
2. *Котов А. А.* Паттерны эмоциональных коммуникативных реакций: проблемы создания корпуса и перенос на компьютерных агентов // Компьютерная лингвистика и интеллектуальные технологии. Вып. 8 (15). — М.: РГГУ, 2009. — С. 211–218.
3. *Котов А. А.* Компенсация коммуникативных стимулов в эмоциональном диалоге // Компьютерная лингвистика и интеллектуальные технологии. Вып. 12 (19). — М.: РГГУ, 2013. — С. 332–341.
4. *Котов А. А.* Коммуникативное поведение при ответе на сложный вопрос в эмоциональном диалоге // Мультимодальная коммуникация: теоретические и эмпирические исследования.– М.: «Буки Веди», 2014. — С. 74–85.
5. *Bänziger T., Scherer K. R.* Using Actor Portrayals to Systematically Study Multimodal Emotion Expression: The GEMEP Corpus // Affective Computing and Intelligent Interaction, LNCS 4738. — Berlin, Heidelberg: Springer-Verlag, 2007. — Pp. 476–487.
6. *Martin J.-C., Devillers L. A.* Multimodal Corpus Approach for the Study of Spontaneous Emotions // Affective Information Processing. — Berlin, Heidelberg: Springer-Verlag, 2009. — Pp 267–291.
7. *Rehm M., André E.* From Annotated Multimodal Corpora to Simulated Human-Like Behaviors // Modeling Communication, LNAI 4930. — Berlin, Heidelberg: Springer-Verlag, 2008. — Pp. 1–17.
8. *Ekman P., Friesen W. V.* Unmasking the face. A guide to recognizing emotions from facial clues. — Englewood Cliffs, NJ: Prentice-Hall, 1975.
9. *Ekman P.* Emotions revealed. Understanding faces and feelings. — W&N, 2003.

References

1. *Bänziger T., Scherer K. R.* (2007), Using Actor Portrayals to Systematically Study Multimodal Emotion Expression: The GEMEP Corpus, Affective Computing and Intelligent Interaction, LNCS 4738, Berlin, Heidelberg, Springer-Verlag, pp. 476–487.
2. *Ekman P.* (2003), Emotions revealed. Understanding faces and feelings, W&N.
3. *Ekman P., Friesen W. V.* (1975), Unmasking the face. A guide to recognizing emotions from facial clues. Englewood Cliffs, NJ: Prentice-Hall.

4. *Grishina E. A.* (2014), Gestures and grammatical characteristics of utterance [Zhesty I grammaticheskie kharakteristiki vyskazyvaniya], Multimodal communication: theoretical and empirical studies [Multimodalnaya kommunicatsyya: teoreticheskiye i empiricheskiye issledovaniya], Moscow, pp. 25–47.
5. *Kotov A. A.* (2009), Patterns of communicative reactions: creation of corpus and transfer to the computer agents [Patterny emotsyonalnykh kommunikativnykh reaktsyy: problemy sozdaniya korpusa i perenos na kompyuternykh agentov], Computer linguistics and intellectual technologies [Kompyuternaya lingvistika i intellektualnye tekhnologii], Issue 8 (15), Moscow, pp. 211–218.
6. *Kotov A. A.* (2013), Compensation of communicative stimuli in the emotional dialogue [Kompensatsyya kommunikativnykh stimulov v emotsyonalnom dialoge], Computer linguistics and intellectual technologies [Kompyuternaya lingvistika i intellektualnye tekhnologii], Issue 12 (19), Moscow, pp. 332–341.
7. *Kotov A. A.* (2014), Communicative behavior during an answer to a complicated question in an emotional dialogue [Kommunicativnoe povedenie pro otvete na slozhnyy vopros v emotsyonalnom dialoge], Multimodal communication: theoretical and empirical studies [Multimodalnaya kommunicatsyya: teoreticheskiye i empiricheskiye issledovaniya], Moscow, pp. 74–85.
8. *Martin J.-C., Devillers L. A.* (2009), Multimodal Corpus Approach for the Study of Spontaneous Emotions, Affective Information Processing, Berlin, Heidelberg, Springer-Verlag, pp. 267–291.
9. *Rehm M., André E.* (2008), From Annotated Multimodal Corpora to Simulated Human-Like Behaviors // Modeling Communication, LNAI 4930, Berlin, Heidelberg: Springer-Verlag, pp. 1–17.

ТЕЛО В ДИАЛОГЕ И НЕКОТОРЫЕ ПРОБЛЕМЫ МУЛЬТИМОДАЛЬНОЙ КОММУНИКАЦИИ: ПРИЗНАК «РАЗМЕР СОМАТИЧЕСКОГО ОБЪЕКТА»

Крейдлин Г. Е. (gekr@iitp.ru),
Хесед Л. А. (lidakhe@yandex.ru)

РГГУ, Москва, Россия

Ключевые слова: мультимодальность, невербальная семиотика, размер, русский язык, семиотическая концептуализация, тело

HUMAN BODY IN AN ORAL DIALOG: THE CORPORAL FEATURE “SIZE OF THE SOMATIC OBJECT”

Kreydlin G. E. (gekr@iitp.ru),
Khesed L. A. (lidakhe@yandex.ru)

RSUH, Moscow, Russia

The paper continues a series of works on multimodal oral communication, many of which were printed in the Proceedings of previous «Dialogs». The interplay of verbal and nonverbal, mainly corporal, Russian sign codes in everyday communication is explored within the framework of the featured approach. The latter is based on the concept and instruments of the semiotic conceptualization of the human body, i.e. the naive map of how Russians think and talk about the body and its parts, organs, corporal liquids, covers, etc. and how Russians use somatic objects in various types of gestures, postures, sign movements, and other body meaningful units.

The core of the semiotic conceptualization holds several sets such as those of somatic objects and their natural language names, the sets of corporal features, their values and names, etc. In this paper, we focus primarily on the series of features named «the size of the somatic object» and provide some results of their language and nonverbal semiotic analyses. Two basic kinds of the features discussed that we call *an absolute size* and *a relative size* are distinguished, and the meaning and usage of many Russian expressions which reflect absolute and relative sized are described. Also, some correlations between some verbal and nonverbal Russian sign units of size are singled out.

Key words: multimodal communication, nonverbal semiotics, size, semiotic conceptualization, Russian, verbal units, nonverbal units

Введение. Постановка задачи

Настоящая работа продолжает серию исследований, направленных на изучение проблемы взаимодействия естественного языка и невербальных знаковых кодов в устном диалоге, известной под названием **проблемы мультимодальности**[1].

Из множества невербальных знаковых кодов, таких как предметный код, акциональный код, пищевой код и др., основное внимание в работе уделяется **телесному,** или **соматическому**, **коду**. К лексическим единицам этого кода относятся, во-первых, слова и словосочетания, обозначающие телесные, или соматические, объекты, признаки таких объектов и значения признаков, а во-вторых, жесты (в широком понимании этого слова). Это знаковые движения рук, ног, плеч и головы (иначе, собственно жесты), а также выражения лица, или мимика, телодвижения, позы и смешанные вербально-невербальные знаковые формы поведения, или манеры.

Изучение разных аспектов мультимодальной коммуникации мы проводим в рамках **признакового подхода**, в основе которого лежит понятие **наивной семиотической концептуализации тела**[2]. Такая концептуализация релятивизована относительно конкретного естественного языка и соответствующего ему языка жестов, каковыми в нашей работе являются русский язык и русский язык жестов.

Русская наивная семиотическая концептуализация тела — это, фактически, модель того, что обычный, то есть не искушенный в науке, русский человек думает и говорит о теле и других телесных объектах, а также того, как он пользуется телом в устных диалогах. Данная модель складывается из большой совокупности телесных объектов, признаков и языковых имён этих объектов и признаков, а также жестов, совершаемых телом или над телом.

***

Одно из центральных мест в системе телесных признаков занимает признак «размер <данного телесного объекта>», прежде всего, потому, что, с одной стороны, этим признаком обладают очень многие **типы соматических объектов**, а с другой стороны, он связан с другими признаками того же самого объекта.

В задачу настоящей работы входит описание признака «размер <данного> соматического объекта», его значений, а также языковых и жестовых единиц, выражающих как сам признак, так и его значения. Попутно мы останавливаемся на отдельных аспектах проблемы мультимодальности, прежде всего, тех,

---

[1] См. подробнее [Крейдлин 2002, 2006, 2007, 2008, 2010] и [Крейдлин, Переверзева 2010, 2013, 2014].

[2] О признаковом подходе и связанных с ним проблемах см., например, [Крейдлин 2010].

которые касаются роли признака размера и его значений в семиотическом представлении отдельных жестов.

Структура дальнейшего изложения такова.

Вслед за Введением, в разделе 1 речь пойдёт о признаке «размер <данного> соматического объекта»[3] и его связи с некоторыми свойствами обладателя этого объекта. Основным содержанием разделов 2 и 3 является определение места признака размера в системе телесных признаков. Раздел 4 посвящён двум основным разновидностям данного признака — **абсолютному** и **относительному размеру**. В разделе 5 рассматриваются разные значения абсолютного и относительного размера и их типовые вербальные выражения. В том же разделе обсуждаются две содержательные характеристики определённых значений признака — **их семантическая** и **культурная выделенность**. В разделе 6 рассматривается место признака «размер» (и его языковых выражений) в мультимодальной коммуникации, то есть его участие в описании отдельных правил взаимодействия языковых выражений значений этого признака с эмблематическими жестами[4]. В Заключении подводятся итоги работы и обозначаются перспективы дальнейших исследований в рамках рассматриваемой проблемы.

1. Признак «размер соматического объекта», его значения и свойства

Описывая человека, люди обычно характеризуют его внешний вид, указывают на рост, особенности телосложения, форму лица, цвет глаз, длину волос и на некоторые другие признаки. Не оставляют они в стороне и размер тела или отдельных его частей. При таком описании характеристики размера часто сочетаются с языковыми выражениями других признаков (причём не только телесных), образуя в совокупности его портрет. Ср. примеры (1)–(3):

(1) *Он* <Печорин> *был* ***среднего роста****; стройный, тонкий стан его и* ***широкие плечи*** *доказывали крепкое сложение, способное переносить все трудности кочевой жизни и перемены климатов* (М. Лермонтов. Герой нашего времени)*;*

(2) *Молодой князь был* ***небольшого роста****, весьма красивый, сухощавый брюнет, с несколько истощённым видом, <…> в чрезвычайно изящной одежде и* ***с крошечными руками и ногами*** (Л. Толстой. Война и мир)*;*

---

[3] В работе используется следующая система условных обозначений: имена признаков телесных объектов выделяются знаком «», значения признаков — знаком //. Имена жестов, а также важные для работы понятия и термины выделяются жирным шрифтом, а языковые выражения и текстовые примеры — курсивом.

[4] О семиотических типов жестов, в частности, об эмблематических жестах, то есть жестах, которые имеют собственное лексическое значение и употребляются в норме без речевого сопровождения, см. подробнее [Крейдлин 2002].

(3) *— Я в одной папиной книге <…> прочла, какая красота должна быть у женщины <…> Тонкий стан,* ***длиннее обыкновенного руки****, — понимаешь, длиннее обыкновенного! —* ***маленькая ножка*** (И. Бунин. Лёгкое дыхание).

В этих примерах выделенные сочетания передают значения таких признаков, как «размер тела» (*среднего роста, небольшого роста*), «размер рук» (*крошечные руки*) и «размер плеч» (*широкие плечи*).

Размером обладает бóльшая часть соматических объектов — тело, его части и части этих частей, телесные покровы (кожа, волосы и ногти), органы, кости, линии (морщины, талия, черты лица) и др. Указывая на размеры разных типов телесных объектов, мы используем такие сочетания, как *крупное тело*, *длинные руки*, *короткие пальцы*, *маленькие ногти*, *длинные волосы*, *увеличенная печень* и *мелкие черты лица*, ср.:

(4) *Черноглазая, с* ***большим ртом****, некрасивая, но живая девочка <…>* (Л. Толстой. Война и мир),

где описывается размер рта девочки, и

(5) *У него* <Лаврентьева> *были пышные волосы и* ***мелкие черты лица*** *— сочетание гнусное* (С. Довлатов. Чемодан),

где указывается размер черт лица персонажа.

Признак «размер <данного> соматического объекта» обладает целым рядом важных свойств.

Во-первых, отдельные его значения могут не только характеризовать конкретные черты внешнего облика, но и указывать на связь телесной системы человека с компонентами других его систем, например психической и ментальной[5]. Об этом говорит пример (6), в котором изменение размера глаз персонажа напрямую связывается с изменением его эмоционального состояния:

(6) *Машина прыгнула. Веня глянул на прокурора <…> И увидел его* ***глаза — большие****, белые* ***от ужаса*** (В. Шукшин. Мой зять украл машину дров!).[6]

---

[5] Об основных системах человека см. [Апресян 1995].

[6] Это свойство размера не является, разумеется, уникальным. Им обладают, например, признаки формы и цвета соматического объекта, ср. предложения:

(7) *Лицо его <Мохнача>, до того величественное и неподвижное, избороздили мелкие морщинки,* ***глаза сузились****, в них заблестели слезы* (П. Алешковский. Рассказы)

и

(8) *У Даши разгорелось лицо, серые* ***глаза потемнели от решимости*** (П. Проскурин. Полуденные сны),

в которых изменения формы и цвета глаз вызваны изменившимся эмоциональным — состоянием героев.

Во-вторых, некоторые значения признака размера данного соматического объекта могут указывать на социальные или культурные характеристики человека — его обладателя, такие, как пол, возраст, расовая, этническая или культурная принадлежность. В диалоге (9) вывод о принадлежности преступника к определённому сообществу делается как раз на основании характеристики его внешнего облика и размера ноги:

(9) *Давайте еще раз вспомним все приметы* <преступника>. ***Маленькая нога*** *<…>, ходит босиком, <…> очень ловок,* ***мал ростом****, отравленные шипы. Какой вы делаете из этого вывод?*
— Дикарь! — воскликнул я. <…>
— Аборигены Андаманских островов могут, пожалуй, претендовать на то, что они самое низкорослое племя на земле (А. Конан-Дойль. Знак четырёх).

За конкретным телесным размером в данной культуре может быть стереотипно закреплена строго определённая функция — служить маркером конкретной характеристики человека. Так, *узкие глаза* могут выдавать принадлежность человека к монголоидной расе, а в качестве стереотипа женской красоты в русской культуре часто выступает *маленькая ножка* (ср. выше пример (3)). *Крупное телосложение* мужчины может говорить о его принадлежности к определённой социальной группе, например, к спортсменам, специализирующимся в определённом виде спорта, а *большие руки* как у мужчин, так и у женщин более свойственны людям, занятым физическим трудом, чем, например, искусством, ср.:

(10) *У неё были усталые* ***большие руки*** *ткачихи или обмотчицы. Такие руки я видел на заводских конвейерах, руки-кормильцы* (Д. Гранин. Месяц вверх ногами).

Ещё одним, третьим, свойством разных соматических объектов является **пропорциональность** или **непропорциональность** их размеров. Например, резкая непропорциональность размеров разных телесных объектов нередко создаёт впечатление о непривлекательной внешности человека, вплоть до уродства. Напротив, подчёркивая пропорциональность размеров, автор текста может специально настраивать читателя на положительное восприятие облика героя. Сказанное иллюстрируют предложения (11) и (12), в которых прилагательные *ладный* и *нескладный* обозначают, соответственно, пропорциональную и непропорциональную фигуры героев:

(11) *В глаза всем сразу бросилась его* <графа> *удивительно* ***ладная фигура****, которая, <…> продолжала сохранять почти юношескую стройность в сочетании с благородной осанкой* (Я. Тирадо. Испанский меч).

(12) *В воспоминаниях Дуни мелькал образ высокой,* ***нескладной девочки****, белобрысой и некрасивой, с умным лицом, такой сухой и чёрствой на вид* (Л. Чарская. Приютки).

В связи с отмеченными свойствами размеров соматических объектов возникает естественный вопрос: а обладают ли этими свойствами объекты произвольной природы? К сожалению, ответ на этот вопрос мы сегодня дать не можем, поскольку системного анализа признака «размер» для объектов разной природы, насколько нам известно, ещё не проводилось.

2. Признак «размер соматического объекта» в системе телесных признаков

Признаки, характеризующие соматические объекты, делятся на три группы: **структурные**, **физические** и **функциональные признаки**[7].

Структурные признаки описывают внутреннее строение и пространственные соотношения данного соматического объекта с другими. К структурным признакам относятся, в частности, (1) возможность членения данного соматического объекта на более мелкие части (так, ладонь и пальцы являются частями кисти); (2) возможность вхождения данного объекта в состав другого объекта (кисть, наряду с плечом и предплечьем, является частью руки) и (3) местоположение и ориентация данного объекта относительно других объектов в составе тела (руки расположены в верхней части тела).

К физическим признакам относятся признаки, которые характеризуют объект сам по себе, вне связи с какими-то другими соматическими объектами. Эти признаки люди обычно воспринимают при помощи разных органов чувств. Например, со зрительным каналом восприятия связан цвет соматического объекта (ср. *рыжие волосы*, *зелёные глаза*, *бледное лицо*); со слуховым каналом — звучание соматического объекта (ср. *кости трещат*, *в животе урчит*), а с тактильным каналом — его текстура, то есть свойства его поверхности (ср. *шершавая кожа*, *гладкие волосы*).

Функциональные признаки бывают двух основных видов — <**собственно**> **функции** и **дисфункции**. Функции указывают на предназначение соматического объекта, обеспечивающее нормальную жизнедеятельность человека — его обладателя, а дисфункции связаны с нарушением нормального функционирования или самого человека, или какого-то его телесного объекта. Кроме того, дисфункциями являются разного рода телесные аномалии, например, отсутствие у человека руки[8].

Что касается «размера соматического объекта», то его нельзя, строго говоря, отнести к какой-то конкретной из трёх групп признаков, поскольку он, как и «форма соматического объекта»[9], является и структурным, и физическим.

---

[7] Об этих группах признаков соматических объектов см. подробно в статье [Крейдлин, Переверзева 2010].

[8] Подробнее о функциях и дисфункциях соматических объектов см. в работе [Аркадьев, Крейдлин: 41–53].

[9] О признаке «форма соматического объекта» см. [Аркадьев, Крейдлин, Летучий 2008: 78–97].

Действительно, с одной стороны, размер — это признак физический, поскольку характеризует геометрические, то есть измеряемые и зрительно наблюдаемые свойства объектов. С другой стороны, он же является структурным признаком, поскольку задаёт внешние характеристики соматических объектов не только самих по себе, но и в сравнении с подобными характеристиками других телесных объектов у того же человека. Например, когда мы говорим о размере глаз (ср. *У него крошечные глазки* или *глазёнки*), мы имплицитно связываем размер глаз с размерами остальных частей лица.

3. Связь признака «размер соматического объекта» с другими телесными признаками

Признак размера тесно связан со многими другими телесными признаками, такими, как «объём» (ср. *полные руки*), «функция» (ср. *длинный язык*), «сила» (ср. *мощный торс*) или «вес» (ср. *тяжёлые кулаки*), но особенно тесно он связан с признаком «форма»[10]. Об этом говорят разного рода языковые единицы, такие как лексические биномы типа *глазки-бусинки* или *глазки-вишенки*, то есть глазки, маленькие по размеру и круглые по форме, ср. (13), атрибутивные конструкции типа *лебединая шея* ‘длинная и узкая’, ср. (14), а также конструкции с выраженным семантическим отношением «часть — целое», в которых «часть» является **заметной** (**салиентной**) деталью, ср. (15):

(13) *Савелий Петрович походил на белую мышь <…>. Узенькое личико <…> с длинным носом, крохотными* ***глазками-бусинками*** *и неожиданно большим ртом* (Д. Донцова. Уха из золотой рыбки)*;*

(14) *Удивительно гармонировало тонкое лицо Инны с изгибом* ***лебединой шеи*** (Г. Гликман. Маэстро Мравинский)*;*

(15) *Клавдия Васильевна: Познакомь, Олег.*
Олег: ***С косой*** *— Вера,* ***с глазами*** *— Фира* (В. Розов. В поисках радости).

О том, что признаки размера и формы телесного объекта тесно связаны друг с другом, свидетельствует, прежде всего, связь между их значениями. Например, если форма объекта вытянута в каком-то измерении, скажем, одно из них намного больше другого (значение признака — /вытянутая форма/), то размер вдоль этого измерения будет доминантным, то есть самым большим из возможных. Действительно, *вытянутое лицо* — это лицо *продолговатое* или *длинное*.

О связи признаков формы и размера говорят также и невербальные знаковые единицы, такие как жесты **сделать большие глаза** и **сделать круглые**

---

[10] О связи признаков размера и формы см., например, в работах [Рахилина 1995: 58–81] и [Iordanskaja, Paperno 1996].

глаза. При лексикографическом описании этих жестов важно отметить, что первый из них указывает на изменение размера глаз, а второй — на изменение формы. Оба изменения происходят, так сказать, в одну сторону, поскольку жесты **сделать большие глаза** и **сделать круглые глаза** приводят к одному и тому же результату, который можно охарактеризовать в равной степени одним из двух способов: (а) *глаза стали большими* и (б) *глаза стали круглыми*. Иными словами, результат изменения размера и формы глаз здесь будет одним и тем же.

В описании этих жестов в явном виде содержится информация о роли признаков размера и формы. Однако связь телесных признаков с жестами не ограничивается указанием их роли в семантическом представлении жестов. Признак размера, о котором мы в основном здесь говорим, участвует в формулировке ряда правил взаимодействия лексических единиц, а именно языковых выражений конкретных значений размера, и невербальных единиц — жестов.

Замечание: Признак размера связан и с другими телесными признаками, однако часто такая связь может быть обнаружена только в результате семантического анализа, причём довольно большой степени глубины. Рассмотрим, например, сочетание *волосатый палец*, значение которого ‘на пальце имеется много волос’, или ‘палец покрыт волосами, причём их много’. Первый компонент значения говорит о текстуре пальца, то есть что на нём есть волосы, а второй — о количестве волос, что волос много. Когда на пальце много волос, он воспринимается зрительно как *объёмный*, а объёмные объекты обычно воспринимаются как *большие*. Отсюда следует, что сочетание *большой волосатый палец* — нормальное, о чём свидетельствует, например, предложение

(16) *Пётр Сергеевич показал* ***большой волосатый палец*** (В. Пелевин. Жёлтая стрела).

В то же время сочетание *маленький волосатый палец* представляется аномальным, во всяком случае, в НКРЯ, материалы которого мы существенно используем, нам не встретилось ни одного предложения с этим сочетанием.

Размер соматического объекта может меняться, и к изменению размера ведут разнообразные процессы. Как показывает (17), изменение размера соматического объекта может быть сопряжено с изменением его внутренней структуры:

(17) *Больные* ***ноги*** *его* <Ломоносова> ***опухали*** (В. Шишков. Емельян Пугачёв).

Такие глаголы, как *опухать*, *надуваться, набухать*, *расширяться*; *сжиматься, скукоживаться*, *съёживаться*, *сморщиваться, усохнуть, высохнуть* и т.п. в своих основных значениях как указывают на изменение внутренней структуры некоего телесного объекта, причём в результате такого изменения размер данного объекта становится больше или меньше обычного. Между тем, в семантической структуре приведённых глаголов не содержится указания

ни на размер, ни на его изменение — информация об изменениях размера, вызванных изменениями во внутренней структуре, извлекается путём глубокого семантического анализа этих слов.

4. Значения признака «размер соматического объекта». Абсолютный и относительный размер

Говоря о значениях признака «размер соматического объекта», важно различать две основные разновидности этого признака: **абсолютный размер** и **относительный размер**.

Абсолютный размер не привязан ни к одной пространственной оси координат, а потому для выражения значений абсолютных размеров телесных объектов в русском языке чаще всего используются слова *большой* и *маленький*. Разумеется, существуют и другие способы обозначения абсолютного размера, причём весьма разнообразные морфологически и синтаксически. Среди таких обозначений есть отдельные слова, например, прилагательные *огромный*, *крошечный*, *громадный*, или существительные *гном*, *лилипут*, *мальчик-с-пальчик*, *формы*[11]. Это также стилистически окрашенные имена *рубильник* и *паяльник* (о большом носе), *вывеска* и *будка* (о большом лице)[12]. Абсолютный размер передают также и различные словосочетания, такие как (а) параметрические конструкции с классификаторами *размер* или *величина*, ср. *руки большого размера*; *исполинского размера* (*голова*); *руки огромной величины*; (б) сравнительные обороты типа *ноги, как у слона*; *голова, как тыква*; *глаза, как тарелки*; *ноги, как спички*[13] или (в) аппозитивные сочетания типа *человек-гора*; *глазки-вишенки*; *ноги-тумбы*.

Относительный размер соматического объекта — это его размер вдоль одной из пространственных осей. Примерами языковых выражений относительного размера являются: (а) прилагательные *длинный*, *узкий*, *высокий*, *приземистый*, *длинноухий*, *коротконогий*, *расширенный*, *удлинённый* и др.; (б) существительные *коротышка*, *карлик*, *великан*, *каланча*, *дядя Стёпа*, а также (в) аппозитивные конструкции *ноги-спички*; *ноги-ходули*; *глаза-щёлочки* и (г) сравнительные обороты типа *нос, как у Буратино*; *зубы, как у лошади*.

---

[11] Это существительное в (форме) Pl. Tant. обозначает большие размеры определённых частей женского тела.

[12] О таких языковых единицах см. [СРС 1997].

[13] В качестве основания сравнения здесь обычно выбирается имя, обозначающее нечто стереотипно большое или стереотипно маленькое, ср.:

(18) *Страшен ты, сосед, ох как страшен <…>. Если бы не* ***уши, как у слона*** *в зоопарке, было бы терпимо* (Ю. Мамлеев. Конец света/Дикая история).

Выбор основания, на котором производится сравнение, может зависеть, однако, не только от свойств самого соматического объекта, но и от культурных, национальных, этнических и др. экстралингвистических предпочтений говорящего или каких-то контекстных условий.

Обратим внимание на то, что характер языкового выражения признака относительного размера нередко зависит от пространственного положения его обладателя, которое чаще всего выражается глаголами стандартного положения тела *стоять*, *сидеть* или *лежать*.

Замечание: Пространственное положение тела вообще играет ключевую роль в описании разных свойств и состояний тела (ср. соотношение слов *ленивый* и *лежебока*, соотношение двух значений слова *колено*, когда одно из значений возникает только при сидячей позе человека, ср. *посадить на колени*[14]), а также действий тела или с телом. Так, обозначение победы над противником может передаваться выражением *положить на обе лопатки*, которое при буквальном прочтении означает ‘сделать так, чтобы соперник лежал перед победителем или под победителем’. Разные действия человека, которые он осуществляет, сопротивляясь действиям другого человека или других людей, передаются глаголами *стоять* или *выстоять* в их переносных значениях. А глагол *работать* в одном из производных значений синонимичен глаголу *сидеть* (тоже в производном значении), ср. *двадцать лет работать на одном предприятии* и *двадцать лет сидеть на одном предприятии*.

Как хорошо известно, рост человека связан с вертикальным измерением тела, иными словами, человек, рост которого измеряют, стоит, а не сидит и не лежит. Когда человек стоит, и мы хотим передать, что у него большой размер тела, то употребляем в норме слово *высокий*, но не **длинный*, ср. сочетание *высокий человек*. Когда же он лежит, мы употребляем слово *длинный*, но не **высокий*, ср. *длинный человек*[15].

Указание на положение тела относительно той или иной пространственной оси, часто вместе с указанием на размер или изменение размера какого-то телесного объекта, входит в физическое описание многих русских **невербальных знаков** — поз, телодвижений или собственно жестов. Так, поза **свернуться в клубок** говорит о том, что человек лежит и при этом размер его тела как бы уменьшается, ср. правильное сочетание *лежать, свернувшись в клубок,* и неправильные **сидеть, свернувшись в клубок* и **стоять, свернувшись в клубок*. Когда говорят, что человек *выпятил живот*, то имеют в виду движение живота вперёд, т.е. вдоль саггитальной оси, и вместе с этим увеличение его размера (реальное или представляемое). А когда говорят, что человек *выставил бедро*, имеют в виду движение бедра вдоль горизонтальной оси, но при этом не имеют в виду, что размер бедра увеличился.

---

[14] См. подробнее в [Крейдлин 2014: 445–462].

[15] Данный пример с незначительными изменениями принадлежит Ю. Д. Апресяну [Апресян 1971: 509–523]. Отметим попутно, что отадъективное имя *длинный* употребляется как кличка и используется по отношению к человеку уже независимо от положения его тела.

5. Семантическая и культурная выделенность значений признака «размер соматического объекта»

Некоторые значения признака «размер соматического объекта», равно как и значения целого ряда других телесных признаков, не только характеризуют сам объект, но и указывают на некоторые свойства его обладателя. Последние могут быть разной природы. К ним, в частности, относятся: (а) пол и возраст человека (так, словом *грудь* может обозначаться грудь мужчины, женщины или ребёнка, однако сочетание *маленькая грудь* в норме применимо только к женщинам); (б) его физическое или психическое состояние (сочетание *расширенные глаза* передаёт чувства ужаса или удивления); (в) профессия или род занятий (ср. нормальное *здоровенные кулаки боксёра* и странное ?*здоровенные кулаки скрипача*); (г) этническая принадлежность (так, естественным образом интерпретируются сочетания *узкие глаза японца* или *еврейский нос* и неестественно — ?*узкие глаза еврея* и ?*японский нос*). Многие из приведённых выражений отражают существующие социальные или культурные стереотипы, относящиеся к телу и его частям.

В большинстве приведённых сочетаний мы имеем дело с явлением **семантической выделенности** конкретного значения признака «размер соматического объекта». Под семантической выделенностью некоторого значения телесного признака в <данной культуре> принято понимать такое его значение, которое **характеризует не только сам телесный объект или его внутренние свойства, но и обладателя этого объекта**[16].

Замечание: Подчеркнём, что, когда мы говорим о семантической (а также о культурной, см. ниже) выделенности значения какого-то признака соматического объекта, мы всегда имеем в виду **семантическую (культурную) выделенность по меньшей мере одного языкового выражения этого значения**. Семантически выделенным является, например, значение /маленький/ у признака «относительный размер лба», которое выражается прилагательным *узколобый*. Оно характеризует человека упрямого и ограниченного, человека недалёкого, не отличающегося широтой интересов, ср.

(19) ***Узколобый**, придирчивый до мелочности, он держал семью в вечном страхе* (Н. Островский. Как закалялась сталь).

Значение некоторого признака соматического объекта может обладать также свойством **«быть культурно выделенным»**[17]. Культурно выделенным мы называем то значение признака соматического объекта, которое, наряду с характеристикой самого объекта, говорит **нечто содержательное о культуре, к которой принадлежит его обладатель**.

---

[16] О понятиях семантической и культурной выделенности см. [Клыгина, Крейдлин 2012: 256–267].

[17] Речь идёт о культурной выделенности по крайней мере одного из значений признака.

Так, по существующим культурным представлениям, закреплённым в русских фольклорных текстах и паремиях, значения *большой лоб* и *маленький лоб* говорят об уме человека, соответственно, о большом и малом. Отсюда понятно, почему в паремии *Лоб что лопата, а ума небогато* содержится противоречие между большим размером лба и недостатком у человека ума, выраженное при помощи союза *а*.

Замечание: Не следует думать, что если значение одного полюса шкалы размера является семантически или культурно выделенным, то таковым будет и значение другого полюса. Например, сочетание *длинный язык* является характеристикой не только языка, но и человека, то есть значение /длинный/ признака «размер языка» является семантически выделенным, а антонимичное ему сочетание *короткий язык* говорит только о размере языка (да к тому же употребляется в основном медиками). Иными словами, значение /короткий/ признака «размер языка» семантически выделенным не является.

6. Признак «размер соматического объекта» и мультимодальная коммуникация

В правила, определяющие физическую реализацию некоторых жестов, входит информация о конкретном значении признака «размер» того телесного объекта, который участвует в рассматриваемой реализации. Например, для осуществления жеста **заплести косу** требуются длинные волосы, а реализация эмоционального жеста удивления **раскрыть глаза** предполагает широко раскрытые глаза.

Некоторые жесты могут различаться по наличию или отсутствию в их стандартной номинации определённого значения признака «размер» некоторого телесного объекта. Например, есть две возможные реализации русского жеста **показать нос**, когда жест исполняется одной рукой, и когда он исполняется двумя руками (см. описание этого жеста в [СЯРЖ 2001]). При второй реализации жест может называться не только *показать нос*, но и *показать длинный нос*; иными словами, в одну из стандартных номинаций этого жеста входит указание на размер.

Наконец, в правила употребления некоторых жестов входит указание на обязательность их сочетания с языковым выражением значения признака «размер некоторого телесного объекта». Мы здесь имеем в виду одно из проявлений смешанного синтаксиса (англ. mixed syntax). В предложения смешанного синтаксиса, а их в русском языке довольно много, в качестве обязательного конструктивного элемента входят жесты, а потому эти предложения являются одним основных объектов, изучаемых в рамках мультимодальной коммуникации. Фразы *Я поймал вот такую рыбу* или *Она вот такого роста* всегда сопровождаются жестами, обозначающими, соответственно, размер рыбы и рост человека — без этих жестов фразы не существуют.

Заключение

В настоящей работе мы подробно проанализировали признак «размер соматического объекта» и отдельные знаковые (языковые и неязыковые) выражения его значений. Нам хотелось продемонстрировать не только (а) разнообразие средств выражения размера, но также (б) особенности соотношения этих средств со знаковыми выражениями значений других телесных признаков и (в) роль этого признака при описании некоторых явлений мультимодальной коммуникации.

Было показано, в частности, что разнообразие средств выражения размера пропорционально степени освоенности соматических объектов и их типов. Например, такие типы соматических объектов, как части тела и части этих частей, в максимальной степени доступны разным каналам восприятия и потому лучше всего прагматически освоены. Не удивительно, что для таких телесных объектов знаковые выражения значений их размера весьма богаты и по морфологической, и по семантической структуре.

Особо были выделены два вида размеров телесных объектов — абсолютный и относительный размер, и было показано, что они часто передаются языковыми единицами вместе с другими признаками.

Рассмотрели мы также и роль признака размера в описании некоторых жестов и в описании некоторых правил образования синтаксических единиц мультимодальной коммуникации. Это правила взаимодействия языковых выражений значения признака размера с некоторыми жестами.

Исследование вербальных и невербальных средств выражения размера соматических объектов может послужить отправной точкой для построения целого ряда содержательных классификаций таких объектов. Такие классификации дадут возможность ответить на многие вопросы о природе и содержании телесной системы человека. Среди них: *Какие соматические объекты русские люди называют большими? Какие телесные объекты характеризуются как крупные, длинные, толстые? Каков набор телесных объектов, которые характеризуются и по размеру, и по форме (по текстуре, по внутренней структуре и т. п.)?*

Однако для ответа на эти вопросы одних классификаций недостаточно. Нужно также описать основные способы выражения размера применительно к соматическим объектам разных типов[18] и проанализировать особенности выражения целого ряда других телесных признаков, тесно связанных с размером[19].

Кроме того, более детальное изучение языковых способов выражения признака размера поможет сформулировать многие правила взаимодействия жестового и речевого кодов в устной коммуникации.

---

[18] Эта работа уже частично проведена. Её результаты мы надеемся опубликовать в коллективной монографии «Образ тела в языке и культуре» (в настоящее время готовится к публикации).

[19] Авторы выражают благодарность А. Б. Летучему за внимание к настоящей работе и ценные замечания.

Литература

1. *Апресян Ю. Д.* (1971). О регулярной многозначности // Известия АН СССР. Отделение литературы и языка. Т. 30. Вып. 6. С. 509–523.
2. *Апресян Ю. Д.* (1995). Образ человека по данным языка: попытка системного описания // Избранные труды. Т. II. Интегральное описание языка и системная лексикография. — М:. Языки русской культуры. С. 348–388.
3. *Аркадьев П. М., Крейдлин Г. Е.* (2010). Части тела и их функции (по данным русского языка и русского языка тела) // Сборник научных работ к 80-летию Ю. Д. Апресяна. — М:. Языки славянских культур. С. 41–53.
4. *Аркадьев П. М., Крейдлин Г. Е., Летучий А. Б.* (2008). Семиотическая концептуализация тела и его частей. I. Признак «форма» // Вопросы языкознания, № 6. С. 78–97.
5. *Клыгина Е. А., Крейдлин Г. Е.* (2012). База данных «Тело и Телесность в языке и культуре (идеология, структура и наполнение)» // Компьютерная лингвистика и интеллектуальные технологии: Труды Международной конференции «Диалог» (Бекасово, 30 мая — 3 июня 2012 г.). Вып 11 (18). — М.: Изд-во РГГУ. С. 256–267.
6. *Крейдлин Г. Е.* (2002). Невербальное поведение людей в деловом общении // Труды Международного семинара «Диалог». (Протвино, 6–11 июня 2002 г.), в 2 т. Т. 1. Теоретические проблемы. — М.: Наука. — С. 227–240.
7. *Крейдлин Г. Е.* (2004) Невербальные гендерные стереотипы: культурно универсальное и культурно специфичное // Языковые значения. Методы исследования и принципы описания (памяти О. Н. Селиверстовой). — М.: МГПУ. С. 133–144.
8. *Крейдлин Г. Е.* (2006). Механизмы взаимодействия невербальных и вербальных единиц в диалоге I. Жестовые ударения // Компьютерная лингвистика и интеллектуальные технологии: Труды Международной конференции «Диалог» (Бекасово, 31 мая — 4 июня 2006 г.). Вып. 5 (12). — М.: Изд-во РГГУ. С. 290–296.
9. *Крейдлин Г. Е.* (2007). Механизмы взаимодействия невербальных и вербальных единиц в диалоге II А. Дейктические жесты и их типы // Компьютерная лингвистика и интеллектуальные технологии: Труды Международной конференции «Диалог» (Бекасово, 30 мая — 3 июня 2007 г.). Вып. 6 (13). — М.: Изд-во РГГУ. С. 300–327.
10. *Крейдлин Г. Е.* (2008) Механизмы взаимодействия невербальных и вербальных единиц в диалоге II Б. Дейктические жесты и речевые акты // Компьютерная лингвистика и интеллектуальные технологии: По материалам Международной конференции «Диалог» (Бекасово, 4–8 июня 2008 г.). Вып. 7 (14). — М.: Изд-во РГГУ. С. 248–253.
11. *Крейдлин Г. Е.* (2010) Тело в диалоге: семиотическая концептуализация тела (итоги проекта). Часть 1: тело и другие соматические объекты. // Компьютерная лингвистика и интеллектуальные технологии. По материалам Международной конференции «Диалог» (Бекасово, 26–30 мая 2010 г.). Вып. 9 (16). — М.: Изд-во РГГУ. С. 230–234.

12. *Крейдлин Г. Е.* (2014) Соматические объекты и некоторые их типы. Проблемы лингвосемиотического описания // Труды Института русского языка им. В. В. Виноградова. 2014, № 2. — М.: 2014. С. 445–462.
13. *Крейдлин Г. Е., Переверзева С. И.* (2010). Тело в диалоге: семиотическая концептуализация тела (итоги проекта). Часть 2: признаки соматических объектов и их значения // Компьютерная лингвистика и интеллектуальные технологии. По материалам Международной конференции «Диалог» (Бекасово, 26–30 мая 2010 г.). Вып. 9 (16). — М.: Изд-во РГГУ. С. 235–240.
14. *Крейдлин Г. Е., Переверзева С. И.* (2013). Тело и его части в разных языках и культурах (итоги научного проекта) // Компьютерная лингвистика и интеллектуальные технологии. По материалам Международной конференции «Диалог» (Бекасово, 29 мая — 2 июня 2013 г.). Вып 12 (19). В 2 т. Т. 1. — М.: Изд-во РГГУ. С. 378–391.
15. *Крейдлин Г. Е., Переверзева С. И.* (2014). Тело в диалоге: ориентация соматических объектов и выражение отношений между людьми // Компьютерная лингвистика и интеллектуальные технологии. По материалам Международной конференции «Диалог» (Бекасово, 4–8 июня 2014 г.) Вып. 13 (20). — М.: Изд-во РГГУ. С. 272–283.
16. *Лебедева Л. Б.* (2000). Семантика «ограничивающих» слов // Логический анализ языка. Языки пространств / отв. ред. Н. Д. Арутюнова, И. Б. Левонтина. — М.: Языки русской культуры. С. 93–97.
17. *Рахилина Е. В.* (1995). Семантика размера // Семиотика и информатика. Вып. 34. С. 58–81
18. *Словарь русского сленга* (1997) / И. Юганов, Ф. Юганова. — М.: Метатекст.
19. *Iordanskaja, L., Paperno, S.* (1996). A Russian — English collocational dictionary of the human body (ed. by R. L. Leed). Columbus, Ohio: Slavica Publishers, Inc http://russian.dmll.cornell.edu/body/accented/index.htm.

References

1. *Apresyan Yu. D.* (1971). On the Regular Polysemy [O regulyarnoy mnogoznachnosti] // Bulletin of the Academy of Science of USSR. Department of Literature and Language [Izvestiya AN SSSR]. Vol. 30. № 6. P. 509–523.
2. *Apresyan Yu. D.* (1995). The Image of a Human Being in the Language: an Essay of Systematic Description [Obraz cheloveka po dannym yazyka: popytka sistemnogo opisaniya] // Selected works [Izbrannye trudy]. Vol. II. Integral Description of Language and System Lexicography [Integral'noe opisanie jazyka i sistemnaya leksikografiya]. — M:. Yazyki russkoy kul’tury. P. 348–388.
3. *Arkad’yev P. M., Kreydlin G. E.* (2010).Body Parts and Their Functions (in Russian Language and in Russian Body Language) [Chasti tela I ih funktsii (po dannym russkogo yazyka i russkogo yazyka tela)] // Collection of Scientific Works to 80s Anniversary of Yu. D. Apresyan [Sbornik nauchnyh rabot k 80-letiyu Yu. D. Apresyana]. — M:. Jazyki slavyanskih kul’tur. P. 41–53.

4. *Arkad'yev P. M., Kreydlin G. E., Letuchiy A. B.* (2008). Semiotic Conceptualization of the Human Body and Body Parts. I. A Feature «Shape» [Semioticheskaya konceptualizatsiya tela I ego chastey. I. Priznak «forma»]. // Voprosy yazyoknaniya, № 6. P. 78–97.
5. *Klygina E. A., Kreydlin G. E.* (2012). The database «Human body and corporeality in natural language and culture (structure, ideology and content)» [Baza dannyh «Telo I telesnost' v jazyke I kul'ture (ideologiya, struktura i napolnenie)»] // Computational Linguistics and Intelligent Technologies. Proceedings of the International Conference «Dialog» [Komp'yuternaya lingvistika i intellektual'nye tehnologii. Trudy Mezhdunarodnogo seminara «Dialog»] (Bekasovo, 30 May — 3 June 2012). Vol. 11 (18). — M.: RSUH publishing house. P. 256–267.
6. *Kreydlin G. E.* (2002). Nonverbal Behavior of People in Business Communication [Neverbal'noe povedénije lyudey v delovom obshchenii] // Proceedings of the International Conference «Dialog» [Trudy Mezhdunarodnogo seminara «Dialog»] (Protvino, 6–11 June 2002), in 2 parts. Part 1. Theoretical Problems [Teoreticheskie problemy]. — M.: Nauka. — P. 227–240.
7. *Kreydlin G. E.* (2004) Nonverbal Gender Stereotypes: Culturally Universal and Culturally Specific [Neverbal'nye gendernye stereotypy: kul'turno universal'noe I kul'turno specifichnoe] // Language Meanings. Methods of Research and Principles of Description (in Memory of O. N. Seliverstova) [Yazykovye zhachenia. Metody issledovania i printsipy opisaniya (pamyati O. N. Seliverstovoy)]. — M.: MGPU. P. 133–144.
8. *Kreydlin G. E.* (2006). Mechanisms of Interaction Between Verbal and Nonverbal Units in the Dialog I. Gesture Accents [Mehanizmy vzaimodeystviya neverbal'nyh i verbal'nyh edinits v dialoge I. Zhestovye udareniya] // Computational Linguistics and Intelligent Technologies. Proceedings of the International Conference «Dialog» [Komp'yuternaya lingvistika i intellektual'nye tehnologii. Trudy Mezhdunarodnogo seminara «Dialog»] (Bekasovo, 31 May — 4 June 2006). Vol. 5 (12). — M.: RSUH publishing house. P. 290–296.
9. *Kreydlin G. E.* (2007). Mechanisms of Interaction Between Verbal and Nonverbal Units in the Dialog II A. Deictic Gestures and Their Types [Mehanizmy vzaimodeystviya neverbal'nyh i verbal'nyh edinits v dialoge II A. Deykticheskie zhesty i ih tipy]. // Computational Linguistics and Intelligent Technologies. Proceedings of the International Conference «Dialog» [Komp'yuternaya lingvistika i intellektual'nye tehnologii. Trudy Mezhdunarodnogo seminara «Dialog»] (Bekasovo, 30 May — 3 June 2007). Vol 6 (13). — M.: RSUH publishing house. P. 300–327.
10. *Kreydlin G. E.* (2008) Mechanisms of Interaction Between Verbal and Nonverbal Units in the Dialog II B. Deictic Gestures and Speech Acts [Mehanizmy vzaimodeystviya neverbal'nyh i verbal'nyh edinits v dialoge II B. Deykticheskie zhesty i rechevye akty]. // Computational Linguistics and Intelligent Technologies. Proceedings of the International Conference «Dialog» [Komp'yuternaya lingvistika i intellektual'nye tehnologii. Trudy Mezhdunarodnogo seminara «Dialog»] (Bekasovo, 4–8 June 2008). Vol. 7 (14). — M.: RSUH publishing house. P. 248–253.

11. *Kreydlin G. E.* (2010) Body in the Dialog: Semiotic Conceptualization of the Body (Results of the Project). Part 1: Body and Other Somatic Objects [Telo v dialoge: semioyicheskaya kontseptualizatsiya tela (itogi proekta). Chast’ 1: telo i drugie somaticheskie ob”ekty]. // Computational Linguistics and Intelligent Technologies. Proceedings of the International Conference «Dialog» [Komp’yuternaya lingvistika i intellektual’nye tehnologii. Trudy Mezhdunarodnogo seminara «Dialog»] (Bekasovo, 26–30 May 2010). Vol. 9 (16). — M.: RSUH publishing house. P. 230–234.
12. *Kreydlin G. E.* (2014). Somatic Objects and Several Their Types. Problem of Linguistic and Semiotic Description [Somaticheskie ob’jekty i nekotorye ih tipy. Problemy lingvosemioticheskogo opisaniya] // Proceedings of the Institute of Russian language named after V. V. Vinogradov. [Trudy Instituta russkogo yazyka im. V. V. Vinogradova]. 2014, №2. — M.: 2014. P. 445–462
13. *Kreydlin G. E., Pereverzeva S. I.* (2010). Body in the Dialog: Semiotic Conceptualization of the Body (Results of the Project). Part 2: Features of the Somatic Objects and Their Meanings [Telo v dialoge: semioyicheskaya kontseptualizatsiya tela (itogi proekta). Chast’ 2: priznaki somaticheskih ob”yektov I ih znacheniya] // Computational Linguistics and Intelligent Technologies. Proceedings of the International Conference «Dialog» [Komp’yuternaya lingvistika i intellektual’nye tehnologii. Trudy Mezhdunarodnogo seminara «Dialog»] (Bekasovo, 26–30 May 2010). Vol. 9 (16). — M.: RSUH publishing house. P.235–240.
14. *Kreydlin G. E., Pereverzeva S. I.* (2013). Human Body and Its Parts in Different Languages and Cultures (Results of the project) [Telo i ego chasti v raznyh yazykah i kul’turah (itogi proekta)] // Computational Linguistics and Intelligent Technologies. Proceedings of the International Conference «Dialog» [Komp’yuternaya lingvistika i intellektual’nye tehnologii. Trudy Mezhdunarodnogo seminara «Dialog»] (Bekasovo, 29 May — 2 June 2013). Vol. 12 (19). In 2 parts. Part1. — M.: RSUH publishing house. P. 378–391.
15. *Kreydlin G. E., Pereverzeva S. I.* (2014). Human Body in a Dialog: The Orientation of Somatic Objects in Its Connection with Human Relations [Telo v dialoge: orientatsiya somaticheskih ob”yektov i vyrazhenie otnosheniy mezhdu lyud’mi] // Computational Linguistics and Intelligent Technologies. Proceedings of the International Conference «Dialog» [Komp’yuternaya lingvistika i intellektual’nye tehnologii. Trudy Mezhdunarodnogo seminara «Dialog»] (Bekasovo, 4–8 June 2014). Vol. 13 (20). — M.: RSUH publishing house. P. 272–283.
16. *Lebedeva L. B.* (2000). Semantics of «Restricting» Words [Semantika «ogranichivayushchih» slov] // Logical Analysis of Language. Languages of Spaces [Logicheskiy analiz yazyka. Yazyki prostranstv] / ed. N. D. Arutyunova, I. B. Levontina. — M.: Yazyki russkoy kul’tury. P. 93–97.
17. *Rahilina E. V.* (1995). Semantics of The Size [Semantika razmera] // Semiotika I informatika. Vol. 34. P. 58–81.
18. *Dictionary of Russian Slang* [Slovar’ russkogo slenga] (1997) / I. Yuganov, F. Yuganova. — M.: Metatekst.
19. *Iordanskaja, L., Paperno, S.* (1996). A Russian — English collocational dictionary of the human body (ed. by R. L. Leed). Columbus, Ohio: Slavica Publishers, Inc. http://russian.dmll.cornell.edu/body/accented/index.htm.

ГЛУБИНА ПРОСОДИЧЕСКИХ ШВОВ В ЗВУЧАЩЕМ ТЕКСТЕ (ЭКСПЕРИМЕНТАЛЬНЫЕ ДАННЫЕ)

Кривнова О. Ф. (okrivnova@mail.ru)

Московский государственный университет
имени М. В. Ломоносова, Москва, Россия

Ключевые слова: фонетика, устная речь, просодическое членение, просодический шов, сегментирующая сила словораздела, паузальный маркер, восприятие, инструментальный анализ

THE DEPTH OF PROSODIC BREAKS IN SPOKEN TEXT (EXPERIMENTAL DATA)

Krivnova O. F. (okrivnova@mail.ru)

Lomonosov Moscow State University, Moscow, Russia

This paper deals with the problem of prosodic phrasing in a spoken text. The introductory section provides a brief description of the background, clarifies basic terms and explains the concept of prosodic break and word boundary strength. The second section contains a short analysis of the current state of research in this area of phrasal prosody, highlights the main directions of the modern fundamental studies and applications, notes their relevance and the need to expand their empirical base. The third section deals with issues related to the local markers of prosodic phrasing, their hierarchy and phonetic means of realization. Here are given the examples of prosodic labeling of poetic and prose texts in the original transcription of famous Russian linguists Scherba and Avanesov with equivalent transcripts using quantitative, graduated scale of prosodic indexes similar to the labeling scheme adopted in foreign prosodic studies. Particular attention is paid to discussion of A.Sanderman's study, which is the most thorough contemporary analysis of prosodic phrasing. The fourth section describes the aim, material, technique and results of of perceptual and instrumental analysis of the location and depth of prosodic breaks carried out by the author of this paper on the Russian material. It is shown that native speakers quite consistently determine the location and depth of prosodic breaks using a 5-point rating scale, but breaks with minimum indexes are clearly opposed to the other types on the probability of their perceptual detection. Correlation of perceptual breaks' evaluation with pause duration at word boundaries is also investigated. In conclusion the material, methods and results of the experimental studies discussed in this paper are compared, the current trends in the use of the data are highlighted, the prospects and challenges for further studies of prosodic phrasing in speech are outlined.

Key words: phonetics, spoken language, prosodic phrasing, prosodic break, word boundary strength, pause marker, perception, instrumental analysis

1. Введение[1]

Даже наивный носитель языка, не имеющий специального лингвистического образования, понимает, что в линейной последовательности слов в тексте соседние слова в разной степени связаны между собой по смыслу, синтаксически и даже фонетически. Этот факт можно интерпретировать как признание разной сегментирующей силы (глубины) словоразделов. В письменном тексте в качестве формальных показателей сегментирующей силы словоразделов (word boundary strength) выступают знаки препинания: их наличие/отсутствие и тип. Знаки препинания не только членят текст на когерентные фрагменты, но и указывают в определенной степени на их относительный иерархический статус. В устной речи аналогичную функцию выполняют просодические средства: паузы, переломы тона и другие фонетические явления на граничных участках соседних слов в последовательности. Фонетическое членение звучащего текста с помощью просодических средств осуществляется говорящим в соответствии с общими принципами фонетической организации речи и с учетом смысловой и синтаксической структуры текста. В русскоязычной литературе для обозначения данного явления используется термин «просодическое (также синтагматическое, интонационное) членение. В англоязычной литературе за просодическим членением (далее ПЧ) прочно закрепился термин «prosodic phrasing», которым мы также будем пользоваться. Просодически маркированные словоразделы между фразовыми просодическими составляющими образуют просодические швы (разрывы) в звучащем тексте, что хорошо отражает англоязычный термин «prosodic break». Логично предположить, что внутренняя иерархия ПЧ на фразовом уровне находит отражение в разной глубине просодических швов (далее ПШ), которая создается использованием разных просодических средств между и на краях фразовых просодических составляющих. В интонационной фонологии многие исследователи разделяют точку зрения, согласно которой иерархический статус просодической составляющей однозначно соответствует глубине ПШ, завершающего эту составляющую. Это положение т. н. строгой поуровневой гипотезы (Strict Layered Hypothesis SLH) разделяется, однако, не всеми интонологами и никогда не проверялось экспериментально на сколько-нибудь представительном речевом материале, см. об этом [Ladd 1986; Ladd, Campbell 1991; Sanderman 1996; Selkirk 1984].

В отечественной лингвистике впервые обратил внимание на ПЧ и его особую природу академик Л. В. Щерба, который писал в частности: «В европейских языках (а вероятно и во многих других) самым могучим средством выражения связи между словами и группами слов является „интонация“, „фразировка“ в самом широком смысле слова» («Восточнолужицкое наречие», Пгр., 1915). Он намного опередил зарубежных коллег как в понимании самого явления, так и в его терминологическом обозначении.

К сожалению, Щерба не занимался подробным изучением ПЧ, но в своих работах он обозначил практически все его отличительные особенности, являющиеся в настоящее время предметом исследования во многих работах

[1] Исследование проведено при поддержке гранта РФФИ 15-06-06103.

по фразовой просодии, однако до сих пор не описаны и не объяснены полностью ни для одного из европейских языков. Это относится, в частности, и к иерархической природе ПЧ. В книге «Фонетика французского языка» Щерба отмечает: «Синтагмы (минимальные единицы интонационного членения) могут объединяться в группы высшего порядка с разными интонациями и в конце концов образуют фразу — законченное целое, которое может состоять из группы синтагм, но может состоять и из одной синтагмы, и которое нормально характеризуется конечным понижением тона» [Щерба 1955]. В этой же книге, наряду с большим количеством французских примеров, приведены авторские транскрипции русского стиха, где используются 4 маркера для ПШ разной глубины: в завершении фоносинтагм |, полуфраз {, фраз |, сверхфразовых единств ||. Отмечена также зависимость ПЧ от стиля и темпа произнесения, т. е. от установки говорящего на степень выразительности речи. Обосновывая необходимость введения в лингвистику речи особого раздела «синтаксической фонетики», Щерба подчеркивал динамическую, деятельностную природу ПЧ как в спонтанной, так и репродуцированной речи (в режиме чтения текста), его глубинную связь с «процессом речи-мысли», с активной грамматикой говорящего. Эту идею развивают в настоящее время в психолингвистике и когнитивной лингвистике.

Мысли, близкие идеям Щербы, можно найти в работах многих русских лингвистов и текстологов первой половины XX в. Так, известный текстолог и стиховед Б. В. Томашевский пишет: «При анализе интонационного строя не следует упускать из виду одну его сторону, которую можно назвать «иерархией» интонации… В живом звучании… от слога мы восходим к слову, а от слова к различным степеням фразового членения, к речевым тактам, фразам, предложениям, периодам… Фразовое членение производится иерархически, с подчинением менее крупных единиц более крупным». Он же, говоря о том, что в прозе «ритм и интонация есть спектр синтаксиса», отмечает статистический характер этой связи, затемненной зависимостью ПЧ от «манеры декламации» говорящего, т. е. опять-таки от навыков производства выразительной речи [Томашевский 1929].

Несмотря на интересное и продуктивное обсуждение иерархической природы ПЧ и контролирующих его факторов в русской текстологической литературе XX в., до 80–90-х годов конкретных исследований ПЧ в речи было очень мало как на материале русского, так и на материале других языков. На это были свои причины: господствующая установка лингвистики на структурную научную парадигму, на анализ письменных текстов, с фокусом на сегментной фонетике, слабые аппаратные возможности фонетики, трудоемкость и сложность исследования просодических явлений в речи.

2. Состояние современных исследований ПЧ в звучащем тексте

В 80–90-е гг. в фонетике произошел «просодический бум», тесно связанный с переходом к функциональной и когнитивной научной парадигме в лингвистике, изучению устного дискурса, «языка в действии», интересом

к компьютерным моделям языка и устной речи, в частности к разработкам по автоматическому синтезу речи, невозможному без понимания функций и природы ПЧ. Все это стимулировало, в свою очередь, создание речевых корпусов с просодической разметкой и разработку компьютерного инструментария для автоматической обработки просодии речевых сигналов.

Отчетливо обозначились и главные направления исследований ПЧ (теоретических, экспериментально-инструментальных, прикладных) с группировкой вокруг следующих проблем:

1. Локальные маркеры (границы) ПЧ — текстовая локализация, глубина создаваемого членения (сегментирующая сила границ, их иерархия), средства фонетической реализации.
2. Квантованная/блочная природа просодических составляющих, их иерархический статус, интегрирующие просодические схемы разного уровня, их фонетическая реализация[2].
3. Функциональный аспект ПЧ, контролирующие факторы: коммуникативные, семантико-синтаксические, психофизиологические (когнитивные, речепроизводящие).

В настоящее время ни по одному из приведенных тематических вопросов нет ответов ни в общей, ни в частной фонетике. Практически все исследователи отмечают недостаточность эмпирической базы, представляющей реальную картину ПЧ в звучащей речи в рамках хотя бы одного типа устного дискурса, в том числе даже в прозаическом тексте при его чтении.

Ниже мы ограничимся рассмотрением вопросов, связанных с локальными маркерами ПЧ (просодическими швами).

3. Локальные маркеры просодического членения: иерархия просодических швов, их фонетическая реализация

Как уже было отмечено выше, идея иерархической природы ПЧ встречается в работах многих русских лингвистов и текстологов первой половины XX в. В частности, Л. В. Щерба в русских транскрипциях фразовой просодии использует 4 граничных маркера с разной глубиной членения: для фоносинтагм |, полуфраз {, интонационных фраз |, сверхфразовых единств ||. Поскольку словоразделы, не маркированные как ПШ, в транскрипциях Щербы никак специально не отмечены, можно считать, что в приводимых им примерах для обозначения сегментирующей силы словоразделов использовалась пятибалльная количественная шкала: 0, 1, 2, 3, 4.

Приведем в качестве иллюстрации фрагмент просодической разметки стихотворения А. С. Пушкина «Памятник» в оригинальном варианте Щербы (1a) и с использованием эквивалентных количественных показателей глубины ПШ (1b) [Щерба 1955: 24]:

---

[2] Заметим, что эффект пограничного сигнала может создаваться просто в точке перехода от одной интегрирующей схемы к последующей.

(1a) 'я -''па-мя-тник-се-бе-во-'здвиг-не-ру-ко-'твор-ный |
кне-'му-не-за-ра-'стёт { на-'ро-дна-я-тро-'па |
во-'знё-сся-'вы-ше-он | гла-'во-ю-не-по-'кор-ной |
а-ле-ксан-'дрийс-ко-го-сто-'лпа ||

(1b) 'я <0>''па-мя-тник <0>се-бе<0>во-'здвиг<0>не-ру-ко-'твор-ный <3>
кне-'му<0>не<0>за-ра-'стёт <2> на-'ро-дна-я<0>тро-'па <3>
во-'знё-сся <0>'вы-ше<0>он <1> гла-'во-ю<0>не-по-'кор-ной <1>
а-ле-ксан-'дрийс-ко-го<0>сто-'лпа <4>

Много текстовых примеров с разметкой ПШ содержится и в книге Р. И. Аванесова «Русское литературное произношение» (1972). Аванесов использует пять особых маркеров, фиксирующих разную глубину ПЧ : -, |, /, //, /// в направлении возрастания плюс чистый пробел, т.е фактически исходит из шестибалльной количественной шкалы. Ниже приводится в качестве примера фрагмент просодической разметки из текста К. Федина «Необыкновенное лето» в оригинальном варианте Аванесова (2a) и с использованием эквивалентных количественных показателей глубины ПШ (2b) [Аванесов, 1972:192]:

(2a) /// Все это водное племя / обладало навыками | долголетних
плаваний // в - большинстве прошло войну / и - самой природой
| было словно выделено | для - пребывания на - судах ///

(2b) <5> Все<1> это <1> водное <1> племя <3> обладало <1> навыками
<2> долголетних <1> плаваний <4> в <0> большинстве <1> прошло
<1> войну <3> и <0> самой <1> природой <2> было<1>словно<1>
выделено <2> для<0>пребывания<1>на <0> судах <5>

Обратимся теперь к практике просодической разметки иноязычных речевых корпусов. Наиболее популярная за рубежом схема Tone and Break Indices (ToBi) предполагает два базовых слоя разметки: тональный и слой макросегментации или, иначе, слой показателей сегментирующей силы словоразделов (Break Indices), которые можно трактовать также как показатели фонетической самостоятельности смежных слов, разделенных словоразделом. После серии тестов на материале английского языка, направленных на поиск такой количественной шкалы для Bi, которая была бы наиболее устойчива к оценкам разных транскрайберов, разработчики ToBi остановились на пятибалльной шкале Bi: 0, 1, 2 ,3 ,4 [Silverman et al. 1992].

По результатам тестов определилось следующее соотношение «брейковых» показателей с иерархией просодических составляющих:

0; 1 — внутри фонетического слова (ФС = PW) и акцентной группы (PW, AccG);

2 — внутри фонетической синтагмы между фонетическими словами/ акцентными группами;

3 — внутри интонационной фразы (IP) между полуфразами/фонетическими синтагмами/акцентными группами;
4 — между интонационными фразами внутри высказывания, а также на границе смежных высказываний.

Наиболее обстоятельный, на наш взгляд, анализ сегментирующей силы словоразделов был осуществлен нидерландской исследовательницей А. Сандерман на материале нидерландского языка [Sanderman 1996]. Автор исследует фонетическую сторону ПЧ, уделяя особое внимание перцептивной оценке локализации и глубины ПШ, для чего ею проведена серия целевых фонетических экспериментов, с последующей статистической обработкой результатов. Корреляция ПЧ с синтаксисом рассматривается в этой работе гораздо менее подробно, да и в целом, исследование проведено на материале т. н. лабораторной речи, т. е. на материале чтения специально отобранных предложений, произнесенных отдельно и в контексте сверхфразового единства. Но даже в таком «усеченном» варианте данная работа является практически единственным примером достаточно полного описания фонетического аспекта ПЧ с учетом его иерархической природы и в связи с разными средствами фонетической реализации и перцептивными оценками.

Приведем наиболее значимые результаты, полученные в исследовании Сандерман. Экспериментально показано, что аудиторы, носители нидерландского языка, дают надежные и согласованные перцептивные оценки сегментирующей силы словоразделов (PBS) в десятибалльной шкале без специальной тренировки или инструкций, в том числе и на делексикализованном речевом материале, что позволяет говорить о том, что оценки даются исключительно на основе доступной слушателям фонетической информации. При этом общая тенденция состоит том, что чем больше просодических маркеров использует говорящий на данном словоразделе, тем более высокие оценки PBS даются словоразделу. Различные комбинации просодических маркеров систематически порождают разные PBS, при этом паузы являются наиболее весомым маркером. Дополнительно обнаружено, что разные дикторы имеют определенные предпочтения в использовании просодических средств для маркирования того или иного словораздела и его глубины. Данные перцептивно-инструментального анализа PBS были далее использованы для тестов на материале синтезированной речи с последующей оценкой ее качества носителями языка. Результаты тестов показали, что просодические маркеры ПЧ структурируют устные высказывания и существенно повышают оценки приемлемости и качества синтезированной речи. При этом наиболее успешным оказался набор правил, моделирующий 5 уровней PBS (из протестированных в диапазоне от 3 до 10). При этом качество синтезированной речи оказалось близким к естественным образцам.

Сандерман отмечает, что ПЧ имеет очевидную коммуникативную функцию, так как контролируется семантико-синтаксическими характеристиками предложений. Однако, даже ограниченный в этом отношении материал ее исследования позволяет сделать вывод, что в просодии гораздо меньше обязательств, чем это предполагается синтаксисом: говорящие имеют большую свободу в использовании просодического маркирования словоразделов и синтаксических границ, причем с весьма ощутимой степенью вариативности.

4. Обнаружение и оценка глубины просодических швов при восприятии звучащего русского текста (экспериментальные данные)

На материале русского языка инструментально-экспериментальных исследований глубины ПШ до середины 90-х г. практически не было: просодическая разметка Щербы и Аванесова, примеры которой приведены выше, основывалась, по-видимому, исключительно на авторской интроспекции. Учитывая это, мы провели исследование, в основе которого лежало предположение, что в звучащем тексте просодическим швам соответствуют такие артикуляторно-перцептивные разрывы (перебои плавности речи) степень обнаружения и локализация которых при восприятии определяется фонетической природой ПШ, их глубиной и смысловой значимостью. Предполагалось также, что носители языка способны достаточно согласованно оценивать степень фонетического разрыва, а тем самым глубину ПШ и сегментирующую силу соответствующего словораздела. Ниже кратко описываются результаты проведенного нами фонетического эксперимента, материал и методика которого в некоторых важных отношениях отличаются от экспериментов Сандерман. Подробнее см. об этом заключение, а также [Кривнова 1995, 1999].

4.1. Материал и методика эксперимента

Наше исследование проводилось на материале русского научного текста лингвистического характера, с достаточно сложным синтаксисом. Текст был прочитан без предварительной подготовки («с листа») диктором, профессиональным лингвистом, носителем московской произносительной нормы, имеющим навыки выразительного чтения текста перед микрофоном (общее время звучания текста около 30 минут, общее количество словоупотреблений 2 215, соответственно словоразделов 2 214).

Мы исходили из того, что естественной мерой перцептивной значимости ПШ является вероятность его обнаружения в воспринимаемом речевом потоке. Можно полагать, что с увеличением глубины членения эта вероятность растет. Для проверки этой гипотезы нами был проведен двухэтапный тест, в котором перед аудиторами-носителями русского языка (10 человек, студентов филологического ф-та МГУ) ставились следующие задачи. На первом этапе испытуемые должны были в режиме реального времени зафиксировать в графическом эквиваленте экспериментального текста, напечатанном без знаков препинания, заглавных букв и абзацного членения, места, где в звучащем эквиваленте текста ощущается разрыв или какое-то нарушение плавности, слитности речи[3]. Текст прослушивался один раз.

---

[3] Отметим, что в прочитанном экспериментальном тексте не было пауз хезитации, поэтому с большой уверенностью можно считать, что все обнаруженные разрывы являются рефлексом ПШ. Очевидно в то же время, что текущие паузальные решения, которые принимаются слушающим в процессе восприятия текста, имеют сложную природу: учитывается вся (т. е. не только фонетическая) текущая информация и внутренние процедурные знания. Никакой специальной делексикализации материала мы не проводили: это весьма сложная и трудоемкая задача.

На втором этапе аудиторы должны были при однократном прослушивании того же текста в режиме реального времени оценить глубину имеющихся в нем разрывов плавности по шкале из 5 баллов (4 соответствует наибольшему разрыву; отсутствие воспринимаемого шва на словоразделе, т. е. 0-степень разрыва, никак специально не маркировалось). На этом этапе задача обнаружения швов в звучащем тексте перед аудиторами не ставилась; каждому из них был дан письменный эквивалент текста, в котором с помощью знака | была указана локализация ПШ, определенная по результатам первого этапа эксперимента: в письменном тексте фиксировались в качестве ПШ только такие разрывы, которые были обнаружены на первом этапе не менее чем двумя испытуемыми. На втором этапе тестирования участвовала та же группа аудиторов, которые прошли через первый этап, временной интервал между этапами — один месяц.

4.2. Результаты эксперимента

При обработке результатов перцептивная оценка глубины ПШ считалась определенной, если большинство аудиторов (75% от общего числа) давали ему одну и ту же оценку. Для неопределенных оценок глубина ПШ определялась как средняя арифметическая всех выставленных аудиторами баллов. Результаты аудиторского анализа и статистические характеристики перцептивных оценок представлены в таблице 1.

Из таблицы видно, что 70% всех ПШ имеют определенные оценки. Следовательно, можно считать оправданным предположение, что разрывы, создаваемые ПШ, в случае их обнаружения слушающими достаточно однозначно оцениваются по глубине членения.

Дополнительно заметим, что подавляющее большинство словоразделов в анализируемом тексте не были вообще отмечены большинством аудиторов как носители ПШ: 1624 из 2214, т. е. 73,3%, а среднее количество просодически немаркированных словоразделов между соседними ПШ оказалось равным примерно 4, а слов соответственно 5, включая служебные.

Таблица 1. Перцептивные оценки глубины ПШ в звучащем тексте

Характер оценки	**Глубина ПШ (в баллах)**				**Всего**	
	1	**2**	**3**	**4**		
	абс. ч-та	**абс. ч-та**	**абс. ч-та**	**абс. ч-та**	**абс. ч-та**	**отн. ч-та (%)**
Определенные оценки	164	80	60	108	412	69, 8
Неопределенные оценки	1,0–2,0	2,0–3,0	3,0–4,0		абс. ч-та	отн. ч-та
	47	87	44		178	30,2
Всего					590	100

Перцептивные оценки глубины ПШ были далее сопоставлены с результатами первого этапа эксперимента. Каждому шву ставились в соответствие

два показателя: глубина и относительная частота перцептивного обнаружения (K / N) × 100 %, где N — общее число аудиторов, K — число аудиторов, обнаруживших данный шов (учитывались только те ПШ, которые получили определенные балльные оценки).

Для каждой интересующей нас перцептивной категории ПШ было построено частотное распределение полученных значений относительной частоты обнаружения и определены средние значения этих распределений. В результате проведенного анализа были получены следующие результаты: средняя частота (вероятность) перцептивного обнаружения ПШ1 = 0,38; ПШ2 = 0,87; ПШ3 = 0,98; ПШ4 = 1,0. Таким образом, исходное предположение о зависимости обнаружения ПШ от его глубины подтверждается.

Обратимся теперь к вопросу, почему ПШ1 характеризуются существенно меньшей вероятностью перцептивного обнаружения. Естественно полагать, что причина этого связана с просодическими средствами реализации ПШ. Из экспериментальнофонетических исследований известно, что в имеющемся наборе просодических маркеров ПЧ наиболее значимым является относительная длительность паузы на словоразделе [Сандерман 1996]. Можно привести также мнение А. Н. Гвоздева, который писал о роли пауз в речи следующее: «По-видимому, в качестве различительного средства фигурирует и длительность пауз; именно разная продолжительность их помогает группировать речевые элементы, устанавливая между ними перспективу по большей или меньшей их близости и намечая единства низшего и высшего порядка. Кажется, нет основания искать устойчиво выраженных, постоянно сохраняющихся степеней длительности пауз; наоборот, они относительны и только соотносятся одна с другой внутри определенного речевого целого» [Гвоздев 1949: 155].

В связи с вышесказанным приведем наши экспериментальные данные.

На рис. 1 показана относительная частота перцептивных оценок глубины ПШ в зависимости от длительности физической паузы на соответствующем словоразделе. Ось Y отражает относительную частоту оценки (в %), ось X — длительность паузы в мс*100. Каждое значение относительной частоты соответствует временному полуинтервалу (100 (n – 1), 100 n), где n = 1, 2, … 10; далее полуинтервалам (1000, 1200), (1200, 1400), (1400, 5000).

Рис. 1 указывает на наличие категориальной границы между ПШ1 и ближайшим к нему по глубине ПШ2 в области пауз длительностью около 200 мс[4]. Таким образом, длительность физической паузы в области 200 мсек является особым показателем для восприятия и оценки перерыва в звучании: здесь происходит качественное и количественное изменение характера перцептивных паузальных решений. Это может интерпретироваться, в свою очередь, как отражение различий в тех условиях, которые сопутствуют появлению этих пауз в порождении речи.

---

[4] Категориальная перцептивная граница, по принятому соглашению, соответствует тому значению переменного физического параметра, которое характеризуется равновероятными перцептивными оценками испытуемых, т. е. в области 50 %.

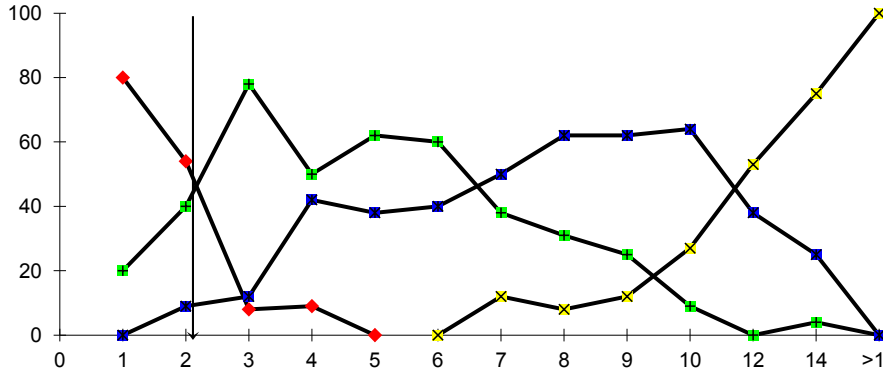


Рис. 1. Относительная частота перцептивных оценок глубины ПШ в % (ось Y) в зависимости от длительности на нем физической паузы в мс*100 (ось X). Параметры кривых: ПШ1 ПШ2 ПШ3 ПШ4 Вертикальная стрелка показывает категориальную границу между ПШ1 и ПШ2.

Интересно отметить, что границам между ПШ большей глубины соответствуют паузальные показатели, которые можно трактовать как целочисленные произведения граничной паузы, типичной для ПШ1 и ПШ2. Так, граница между ПШ2 и ПШ3 обнаруживается в интервале 600–700 мс, а для ПШ3 и ПШ4 в интервале 1100–1200 мс. Если учесть, кроме того, что 200 мс — это средняя длительность слога при нормальном темпе произнесения, можно сказать, что паузы на ПШ глубины 1, 2, 3, 4 по длительности эквивалентны примерно 1, 3, 6 и 7 слогам соответственно.

Последняя, слоговая, мера глубины/сегментирующей силы ПШ и соответствующего словораздела удобна для сопоставления инструментальных данных речи говорящих с разным индивидуальным темпом произнесения.

Из рис.1 также видно, что длительность паузы на ПШ одинаковой глубины варьирует достаточно сильно, даже в области наиболее вероятных оценок.

Физическая пауза, однако, не является единственным локальным маркером ПШ, большую роль играют также фонетические явления на граничных/краевых участках просодических составляющих (т. н. edge-effects). Интонологи указывают на особую просодическую активность *терминальной* зоны, зоны каденций [Николаева 1987], а известный типолог Дж. Гринберг, отмечая универсальное предпочтение языков к особому маркированию *концов* значимых единиц, а не их начал, пишет:«Это, вероятно, связано с тем фактом, что мы всегда можем узнать, когда некто начал говорить, но, как свидетельствует наш печальный опыт, без определенного показателя мы не сможем узнать, когда же говорящий ‘кончит говорить» [Гринберг 1970]. Вопрос о перцептивной значимости и вкладе отдельных просодических параметров в реализацию ПШ разной глубины и сегментирующей силы требует отдельного исследования.

5. Заключение

Подводя итоги описанного выше перцептивного эксперимента в сопоставлении с аналогичными, упомянутыми выше работами зарубежных исследователей, сделаем в заключение ряд дополнительных замечаний.

1. Результаты целевых экспериментов по оценке сегментирующей силы словоразделов на материале разных языков (английского, нидерландского и русского) свидетельствуют о поведенческой устойчивости перцептивных оценок, полученных с использованием иерархической шкалы, различающей не более 5 уровней глубины членения. Это позволяет предположить, что данная шкала отражает какие-то универсальные свойства перцептивных ощущений человека в анализируемом фонетическом пространстве. Для многих практических задач достаточно учитывать этот результат и даже сознательно использовать более крупную трехуровневую шкалу, как это советует делать А. Сандерман на начальной стадии технологических разработок в области синтеза и распознавания речи. Однако для подтверждения указанной гипотезы с общефонетических научных позиций необходимо увеличить как количество и разнообразие рассматриваемых языков, так и число испытуемых, привлекаемых к ее верификации.

2. Для будущих исследований важно также учитывать определенные методические нюансы, проявившиеся даже в тех немногочисленных исследованиях, о которых шла речь в настоящей работе. Прежде всего, возникает вопрос о речевом материале, на базе которого проводится исследование глубины ПШ и иерархии ПЧ в целом. Наш эксперимент, например, проводился на материале естественного связного текста с достаточно сложным и разнообразным синтаксисом, который был озвучен непрофессиональным диктором. В экспериментах Сандерман, как уже было отмечено выше, использовался список предложений, озвученных отдельно или в небольшом контексте как профессиональным, так и непрофессиональными дикторами. Синтаксис этих предложений контролировался экспериментатором, причем могли использоваться и специально сконструированные «искусственные» примеры. В экспериментах с использованием ToBi-разметки на материале английского языка преобладают небольшие новостные тексты, озвученные профессиональными дикторами или радиоведущими. Каждый тип материала имеет свои преимущества и сложности, которые мы не можем здесь обсуждать, однако, очевидно, что выбор того или иного материала требует экспликации и обоснования. Есть различия и в процедуре получения перцептивных оценок глубины словоразделов и ПШ. Так, в зарубежных работах предполагается оценка сегментирующей силы *каждого* словораздела в тексте, причем в экспериментах Сандерман испытуемые (носители языка) дают оценки исключительно на основе своих перцептивных ощущений при многократном прослушивании материала, в то время как в ToBi-экспериментах оценки брейковых показателей получены от экспертов, которым разрешалось использовать не только собственные перцептивные ощущения, но и сопутствующие акустические данные. Наш эксперимент основывался только на перцептивных ощущениях, возникающих в реальном времени, и в то же время двухэтапная процедура получения перцептивных оценок, не предполагавшая оценки

глубины членения для каждого словораздела, значительно облегчала, на наш взгляд, задачу испытуемых и адекватность их окончательных оценок. В целом, и здесь требуется обсуждение возможных процедурных стратегий и поиск оптимальной для каждой конкретной задачи.

3. Экспериментальное исследование, положенное в основу настоящего доклада, исходно имело двоякую цель: во-первых, проверить экспериментально идеи, связанные с иерархической природой ПЧ и ее отображением в глубине ПШ; во-вторых, получить (с учетом имеющихся технических возможностей) речевой материал достаточного объема с антропоморфно адекватной разметкой глубины ПШ, который мог бы далее использоваться для разноаспектных исследований ПЧ в русском звучащем тексте. По нашему мнению, эта цель была в значительной мере достигнута: на полученном материале была проведена большая серия экспериментальных исследований, результаты которых дополнительно подтвердили гипотезу иерархической природы ПЧ и обобщены в докторской диссертации «Ритмизация и интонационное членение текста в «процессе речи-мысли» (опыт теоретико-экспериментального исследования)», защищенной нами в 2007 году. Здесь нет возможности и смысла рассматривать сколько-нибудь подробно даже основные положения этой работы, отметим лишь, что в диссертации была сделана попытка включить феномен ПЧ с учетом его иерархической природы в динамическую модель формирования звуковой стороны речевого высказывания в процессе его порождения и озвучивания, в том числе в составе действующего синтезатора русской речи, иначе говоря, в рамках «активной грамматики говорящего» по Щербе. Кроме того, экспериментально была доказана возможность перцептивной сегментации звучащего текста на предложения–фразы на основе просодической информации без использования лексико-синтаксических текстовых ключей. Были также проведены эксперименты, направленные на уточнение и конкретизацию роли ПЧ в смысловом анализе текста при его восприятии.

4. Несмотря на прогресс, достигнутый в изучении просодического членения на материале разных языков в последнее время, многие вопросы продолжают оставаться открытыми, в том числе и в отношении иерархической организации ПЧ. Так, необходимо дальнейшее исследование факторов, контролирующих текстовую локализацию и глубину ПШ, а также анализ фонетических средств, которые обеспечивают их физическую реализацию. Многие исследователи считают, что решение этих задач возможно только при наличии представительных звуковых корпусов, снабженных антропоморфно адекватной просодической разметкой, а также частеречными и синтаксическими аннотациями произносимого текста. В особенности это актуально для прикладных разработок в области синтеза и распознавания речи, а также автоматизации создания самих звуковых корпусов. В аудитории конференции «Диалог» нет необходимости специально доказывать, что создание таких корпусов требует много временных, кадровых и технологических ресурсов, коллективного участия разных специалистов в области фундаментальной и компьютерной лингвистики. Здесь уместно сказать, что речевая группа кафедры теоретической и прикладной лингвистики МГУ им. М. В. Ломоносова при участии сотрудников некоторых других научных и учебных организаций начинает в этом году работу над проектом «Иерархия

просодического членения звучащей речи: контролирующие факторы и средства реализации» по гранту РФФИ. В ходе работы над этим проектом мы надеемся получить новые результаты, которые будут способствовать более глубокому пониманию просодической организации и членения звучащей речи.

Литература

1. *Аванесов Р. И.* (1972) Русское литературное произношение. Просвещение, М.
2. *Гвоздев А. Н.* (1949) О фонологических средствах русского языка. М.-Л.
3. *Гринберг Дж.* (1970) Некоторые грамматические универсалии, преимущественно касающиеся порядка значимых элементов // Новое в лингвистике. В. 5. М.
4. *Кривнова О. Ф.* (1995) Перцептивная и смысловая значимость просодических швов в связном тексте // Проблемы фонетики, В. 2. Наука, М., сс. 229–238.
5. *Кривнова О. Ф.* (1999) Смысловая значимость просодических швов в тексте // Проблемы фонетики, В. III. Наука, М., сс. 247–257.
6. *Николаева Т. М.* (1989) Три типа высказываний и иерархия интонационной нагруженности // Бюллетень Фонетического Фонда Русского Языка. N 2. Bochum-Ленинград, сс. 8–10.
7. *Томашевский Б. В.* (1929) Ритм прозы // О стихе. Л.
8. *Щерба Л. В.* (1955) Фонетика французского языка. М.

References

1. *Avanesov R. I.* (1972) Russian Literary Pronunciation [Russkoe literaturnoe proiznoshenie], Education, M.
2. *Gvozdev A. N.* (1949) About Phonological Means of the Russian Language [O phonologicheskih sredstvah russkogo jazyka], M.-L.
3. *Grynberg G.* (1970) Some universals of grammar, mainly concerning the order of significant elements [Nekotorye grammaticheskie universalii, preimushchestvenno kasajushchiesja porjadka znachimyh elementov] // The New in Linguistics, V. 5, Foreign Literature, Moscow.
4. *Krivnova O. F.* (1995) Perception and semantic relevance of prosodic breaks in spoken text [Pertseptivnaja i smyslovaja znachimost prosodicheskih shvov v svjaznom tekste] // Problems of Phonetics, V. 2, Science, Moscow, pp. 229–238.
5. *Krivnova O. F.* (1999) Semantic significance of prosodic breaks in spoken text [Smyslovaja znachimost prosodicheskih shvov v svjaznom tekste] // Problems of Phonetics, V. III, Science, Moscow, pp. 247–257.
6. *Ladd R.* (1986) Prosodic phrasing: a case of recursive prosodic structure. Phonology Yearbook 3, pp. 311–340.
7. *Ladd B., Campbell D. R.* (1991) Theories of prosodic structure: evidence from syllable duration // Proc. of the 12th Congress of Phonetic Sciences, Aix-en-Provance, France, pp. 290–293.

8. *Nikolaeva T. M.* (1989) The three types of utterances and the hierarchy of intonation loading [Tri tipa vyskazyvanij I ierarhija intonatsionnoj nagruzhennosti] // Bulletin of the Phonetical Fund of the Russian Language, N 2. Bochum-Leningrad, pp. 8–10.
9. *Sanderman A.* (1996) Prosodic Phrasing (production, perception, acceptability and comprehension). Eindhoven.
10. *Selkirk E.* (1984) Phonology and syntax: the relation between sound and structure, MIT, Cambridge.
11. *Silverman K., Beckman M., Pitrelli J., Ostendorf M., Wightman C., Price P., Pierrehumbert J., Hirshberg J.* (1992) ToBi: a standard for labeling English prosody // Proc. of the International Conference on Spoken Language Processing (ICSLP), pp. 867–870.
12. *Tomashevskij B. V.* (1929) The rhythm of a prose [Ritm prozy] // About poetry, Leningrad.
13. *Shcherba L. V.* (1955) Phonetics of the French language [Phonetika frantsuzskogo jazyka] // Publishing house of foreign literature, Moscow.

ЧАСТИЦЫ «ВОТ» И «ВОН»: МЕХАНИЗМЫ ФОРМИРОВАНИЯ ПЕРЕНОСНЫХ ЗНАЧЕНИЙ НА ОСНОВЕ ИСХОДНЫХ ДЕЙКТИЧЕСКИХ ЗНАЧЕНИЙ[1]

Крылова Т. В. (ta-kr@yandex.ru)

ИРЯ РАН им. В. В. Виноградова, Москва, Россия

Ключевые слова: семантика, дейксис, указательные частицы, метафорические значения

PARTICLES 'VOT' AND 'VON': THE MECHANISMS OF SECONDARY MEANINGS FORMATION ON THE BASIS OF DEICTIC VALUES

Krylova T. V. (ta-kr@yandex.ru)

V. V. Vinogradov Russian Language Institute of the Russian Academy of Sciences

This article is devoted to consideration of the particles VOT and VON. It is another attempt to bring secondary meanings of VOT and VON from their deictic meanings. After analyzing derived meaning of these particles we found no parallelism in the structure of their polysemy. We suggested that this is due to differences in the localization of the object in their index meaning ('proximity to the speaker' VS. 'remoteness from the speaker'). Next, we made an attempt to trace how these components are transformed in a secondary meanings of these words. We found that the component 'proximity to the speaker', which is included in the sense of VON, is transformed into component 'proximity to the moment of speech'. The last one passes in its transformation following stages (each the next is characterized by increasing of metaphoricalness): 'proximity to the moment of speech' → 'temporal proximity of the events' → 'interdependency of events'.

[1] Работа выполнена при поддержке программы фундаментальных исследований Президиума РАН «Историческая память и российская идентичность» («Основной лексический фонд русского языка как элемент русской культуры: системная организация лексики и ее отражение в словаре»), гранта НШ-3899.2014.6 для поддержки научных исследований, проводимых ведущими научными школами РФ («Разработка материалов для активного словаря русского языка»), а также гранта РГНФ № 13-04-00307а «Подготовка второго выпуска Активного словаря русского языка».

Meanwhile the component ‘remoteness from the speaker’ included in the meaning of VON is transformed into component ‘high degree’ or into the indication of a reference to some fact (commonly known to speaker). Thus we can conclude that the asymmetry of VOT and VON derived values is due to the different direction of metaphorization of deictic components ‘proximity to the speaker’ VS. ‘remoteness from the speaker’.

Key words: semantics, deixis, index particles, derived meanings

1. Введение

Частицы ВОТ и ВОН неоднократно становились объектами лингвистического рассмотрения. В частности, они рассматриваются в статье [Борисова, Овчинникова 2011], где ставится задача, близкая к нашей, а именно, изучение метафоризации пространственных отношений, обозначаемых этими частицами.

Авторы этой статьи акцентируют внимание на отсутствии параллелизма а семантике дейктических значений ВОТ и ВОН; с их точки зрения, эти частицы различаются не столько различием в локализации объекта, сколько типом указания: ВОТ отождествляет предмет (лицо, место и даже действие или качество) с наименованием, тогда как ВОН задает способ нахождения нужного объекта.

Мы не можем согласиться с таким описанием различий между этими словами, хотя бы потому, что во многих контекстах свести значение ВОТ к отождествлению едва ли возможно: ср. контексты типа *Мне нравится вот это платье; Делайте вот так; Вот, пожалуйста; Вот, возьмите; Вот тут я живу; Поставьте кресло вот сюда; Вот сколько у нас еды!; Вот какой он большой!*

Аналогичным образом, понять ВОН как указание на способ нахождения предмета можно, пожалуй, только в контекстах типа *Вон там я живу;* для большинства других контекстов (*Мне нравится вон то платье; Вон идет Иван*) такое понимание значения ВОН едва ли можно признать правомерным. С другой стороны, функция идентификации для ВОН типична почти в такой же степени, как и для ВОТ; ср. *Вон моя школа; Вон та девушка — моя сестра.*

С нашей точки зрения, ВОТ и ВОН в своем основном, указательном, значении демонстрируют семантический параллелизм, что отражается в единой схеме построения их толкований.

ВОТ 1.1 ‘Говорящий[2] показывает адресату объект, ситуацию или место А1, причем А1 находится или происходит в том же пространстве, в котором говорящий мыслит себя’.

Ср. *Вот моя квартира; Вот это моя мама; Дайте мне вот это пирожное; Смотри, вот они садятся за стол.*

---

[2] В данном толковании должно также присутствовать указание на речевой акт: ‘говорящий сообщает, что в момент речи он показывает объект…’. В дальнейшем мы будем опускать компонент ‘сообщает’ в целях упрощения толкований.

ВОН 1.1 ‘Говорящий показывает адресату объект, ситуацию или место А1, причем А1 обычно находится или происходит в пространстве, удаленном от того, в котором говорящий мыслит себя’. Ср. *Вон мой дом; Вон та кукла самая красивая; Вон бежит собака*.

Как мы видим, толкования ВОТ 1.1 и ВОН 1.1 практически идентичны; их отличает только характер пространства, в котором находится объект. Именно из этого различия, на наш взгляд, вырастают различия в метафорических значениях ВОТ и ВОН.

Оба рассматриваемых слова имеют развитую многозначность, при этом по количеству значений ВОТ в 3 раза превосходит ВОН. (У ВОТ мы выделили 17 значений, объединяемых 10 более крупных блоков, у ВОН — 6 значений, объединяемых в 4 блока).

Перечислим (максимально схематично) основные значения этих двух слов.

ВОТ

вот 1.1 ‘говорящий что-то показывает адресату’: *Вот моя квартира*.

вот 1.2 ‘говорящий сообщает адресату ответ на какой-л. вопрос’: *Вот что мы сделаем — мы ему обо всем расскажем сами*.

вот 2 ‘говорящий предлагает адресату представить себе что-л.’: *Вот пришел он к ней и говорит: «Выходи за меня замуж»*.

вот 3.1 ‘говорящий сообщает, что время события близко к моменту речи’: *Вот только что <сию минуту> я его видел, и уже опять он куда-то исчез*.

вот 3.2 ‘говорящий подчеркивает, что какое-то событие произошло незадолго до момента речи, причем его результаты сохраняются в момент речи’: *Вот и лето пришло*.

вот 3.3 ‘говорящий сообщает актуальную информацию о себе:’ — *Ты откуда? — Да вот приехал отца навестить*.

вот 4 ‘говорящий подчеркивает, что желательное событие произойдет сразу после другого события’:*Вот сделаю уроки и пойду гулять*.

вот 5 ‘говорящий угрожает адресату’: *Вот скажу маме, как ты меня обзываешь, она тебе задаст!*

вот 6.1 ‘говорящий выделяет кого-л. или что-л.’: *Вот вы, к примеру, что об этом думаете?*

вот 6.2 ‘говорящий подчеркивает, что он переходит от общих утверждений к конкретным примерам’: *Так вести себя некрасиво. Вот скажи, разве тебе было бы приятно, если бы тебя стали дразнить плаксой?*

вот 7 ‘поэтому’: *Он сломал велосипед, вот пусть теперь и чинит его*.

вот 8.1 ‘говорящий дает эмоциональную оценку какого-л. объекта или ситуации’: *Вот сумасшедший!*

вот 8.2 ‘говорящий подчеркивает несогласие с адресатом’: *А вот и ошибаешься*.

вот 8.3 ‘говорящий выражает желание, чтобы имела место какая-то ситуация’: *Вот бы оказаться сейчас на берегу моря*.

вот 8.4 ‘говорящий подчеркивает свое одобрение действий адресата’: *Вот, правильно*.

вот 9 ‘говорящий сообщает, что он закончил высказывание’: *Я считаю, что вы поступаете неправильно. Вот.*

вот 10 ‘говорящий передает чужие слова’: *Мне говорят, вот, я плохая мать.*

ВОН

вон 1.1 ‘Говорящий что-то показывает адресату’: *Вон мой дом.*

вон 1.2 ‘Говорящий указывает на высокую степень чего-л.’: *Наверное, они опять поругались — вон какой он мрачный.*

вон 2.1 обиходн. ‘Говорящий сообщает, что он понял суть дела’: *Ах вон в чем дело — она передумала продавать дом.*

вон 2.2 ‘Говорящий ссылается на плохую ситуацию, известную адресату’: *Я с тобой по- хорошему, а ты вон что делаешь!*

вон 3, разг. ‘Говорящий иллюстрирует свое утверждение’: *Совсем дети распоясались. Вон, мой младший вчера бабушку до слез довел.*

вон 4, обиходн. ‘Говорящий выделяет кого-то или что-то’: *Мне вон никто не помогал.*

Поскольку в своем основном значении ВОТ и ВОН обнаруживают значительное сходство, можно ожидать, что похожей будет и структура их полисемии (набор значений, принципы их объединения в блоки). Тем не менее, оказывается, что это не так. Среди производных, недейктических значений ВОТ и ВОН, по сути дела, есть только два значения[3], которые сближаются между собой.

2. Метафорические значения *вот* и *вон* со сходной семантикой

2.1. Ответ на вопрос

ВОТ 1.2 ‘Говорящий формулирует для себя или сообщает адресату ответ А2 на вопрос А1, заданный кем-то или возникший у самого говорящего’. Ср. примеры:

(1) *Ах, вот зачем она меня позвала — чтобы отчитать за вчерашнее.*

(2) *Я придумал, вот что мы сделаем — мы ему обо всем расскажем сами.*

(3) *Так вот за кого она выходит замуж — за брата Пети.*

---

[3] В дальнейшем мы рассмотрим еще одно (иллюстративное) значение, в котором ВОТ и ВОН частично сближаются, хотя между ними сохраняются существенные различия. Ср. *Всякое отступление от линии партии — это смерть или предательство. Вон, какие люди были, а как скатились в болото оппозиции,* [...] *то вон к чему и пришли!* (Ю. Домбровский); в этом примере возможна замена ВОН на ВОТ. О различиях между ВОТ и ВОН в соответствующем значении см. в разделе 4.

ВОН 2.1 ‘Говорящий формулирует ответ А2 на вопрос А1, обычно возникший у него самого, причем А2 является неожиданным для говорящего’.

(4) *Ах вон в чем дело — она передумала продавать дом.*

(5) *Вон ты чего хочешь — поехать на юг без меня.*

Рассмотренные лексемы сближает еще и то, что обе они могут употребляться, когда А2 не выражено; в этом случае их значение модифицируется следующим образом: ‘Говорящий, который долгое время хотел знать ответ на вопрос А1, узнал или понял, каков он’: *Вон <вот> ты о чем!; Вон <вот> ты где!*

Семантическая связь между основным, указательным значением и рассматриваемым значением является вполне прозрачной: смысл ‘говорящий делает так, чтоб адресат видел объект А1’ (= ‘показывает адресату А1’) модифицируется в смысл ‘говорящий делает так, чтобы адресат знал ответ А2 на вопрос А1’ (= ‘сообщает адресату А2’).

Различие в пространственной локализации объекта, которое фиксируется между ВОТ 1.1 и ВОН 1.1 (близость к говорящему VS. удаленность от него), в значениях ВОТ 1.2 и ВОН 2.1 переосмысляется как различие в характере информации: дейктический компонент ‘удаленность от говорящего’ трансформируется в компонент «неожиданность для говорящего» в случае ВОН 2.1.

Действительно, как видно из вышеприведенного толкования, ВОН в описываемом значении всегда предполагает, что говорящий получил ответ на вопрос, который волновал его самого, причем узнал что-то неожиданное для себя. Ср. следующие типичные примеры:

(6) *Мужик оглянулся и, улыбаясь, сказал вполголоса: — А, это вон кто! Помнишь, вместе на мельнице были?* (Г. Марков).

(7) *Э, вон для чего тебя сюда муж прислал: выведать у меня о Сухомлине!* (В. Василевский).

(8) *А, вы вон про что! — наконец раскусил дядя* (Ю. Домбровский).

В значении ВОТ 1.2 компонент «неожиданность для говорящего» отсутствует.

Во-первых, это слово может использоваться не только тогда, когда говорящий отвечает на свой собственный вопрос, но и тогда, когда он сообщает адресату ответ на вопрос, который тот задал или мог бы задать; ср. *Вот что я вам посоветую: обратитесь к гомеопату.*

Во-вторых, ВОТ 1.2 часто употребляется, когда говорящий строит высказывание в форме ответа на вопрос только для того, чтобы сделать акцент на какой-л. важной информации. См.

(9) *Обгорелые доски да груда кирпичей — вот и все, что осталось от их дома; Живость и непосредственность — вот что меня в ней привлекало.*

(10) *Не опаздывай — вот единственное, о чем я тебя прошу. Во всех этих контекстах употребление ВОН было бы невозможно.*

В-третьих, даже в контекстах, в которых говорящий формулирует ответ на свой собственный вопрос, ВОТ 1.2 не обязательно предполагает обнаружение какой-л. новой и неожиданной для говорящего информации. Эта лексема может использоваться не только в ситуации, когда субъект получил информацию «со стороны», узнав новые факты или сделав выводы на их основании, но и когда он отвечает на какой-л. поставленный им самим вопрос, не пользуясь «внешней» информацией — в частности, вспоминает забытые факты, подыскивает в уме объект с нужными свойствами, анализирует уже имеющуюся информацию и пр. Ср.

(11) *Я вспомнил: Кривой рог — вот как назывался тот город, куда я ездил в детстве.*

(12) *Меня осенило: бинокль — вот что нужно ему купить.*

(13) *Я понял: Иван — вот кто мог это сделать. Во всех этих контекстах использование ВОН было бы неуместно, поскольку это слово в соответствующем значении предполагает, что говорящий узнал что-то новое для себя.*

2.2. Выделение

ВОТ 6.1 = ВОН 4 ‘Говорящий выделяет конкретный объект А1 внутри класса объектов, чтобы узнать или сообщить о нем А2’.

Примеры употребления ВОТ 6.1:

(14) *Вот вы, к примеру, что предпочитаете на обед?* (А. Дорофеев).

(15) *Немало наездился на повозках, санях, верхом на лошадях в детстве, потом на автомобилях.* […] *А вот летать на самолёте ещё не приходилось* (В. Быков).

(16) *Вот котелок жалко, вещь, необходимая всюду, в пути в особенности* (В. Астафьев).

(17) *А вот дядюшек он, напротив, любил, к каждому из них его по-своему влекло* (А. Варламов).

Примеры употребления ВОН 4:

(18) *Вон Саша все успевает.*

(19) *Как там мой дедок? Надо, надо подлечить его. Обязательно на ноги поставить. Они вон только двое у меня со старушкой* (В. Валеева).

(20) *Вон пенсионеры всякие знают, куда в таких случаях обращаться* (В. Войнович).

(21) *Ты же талантливая. Меня вон, как лоха, развела* (А. Геласимов).

Данное значение ВОТ и ВОН, опять-таки, выводится из основного, указательного, при этом выстраивается следующая смысловая цепочка: ‘говорящий делает так, чтоб адресат видел объект А1’ → ‘говорящий привлекает внимание адресата к объекту А1’ → ‘говорящий выделяет объект А1 среди однородных элементов’.

Несмотря на значительное смысловое сходство, между ВОТ 6.1 и ВОН 4 фиксируется ряд различий. Вторая из них значительно менее свободно употребляется в рассматриваемом значении, что проявляется в наличии ряда стилистических, сочетаемостных и синтаксических ограничений на ее использование[4]. Возможно, это связано с тем, что выделение объекта из группы однородных более естественно в том случае, когда объект находится близко от говорящего.

Остальные недейктические значения ВОТ и ВОН не обнаруживают параллелизма и не соотносятся друг с другом.

Как нам кажется, это связано с уже отмеченным выше различием в пространственной локализации объекта, которое фиксируется у описываемых частиц в их основном, дейктическом значении. Именно различие в пространственной локализации в рамках указательного значения оказывается причиной принципиального расхождения в составе производных значений и структуре полисемии ВОТ и ВОН.

3. Частица *вот*. Метафоризация идеи пространственной близости

Для ВОТ актуально несколько центральных семантических идей, вокруг которых группируются его производные значения и которые, в свою очередь, тесно связаны между собой. Все они возникают в результате метафорического переосмысления идеи пространственной близости. Мы считаем целесообразным выделить 3 таких идеи.

3.1. Близость к моменту речи

Первостепенное значение для семантики метафорических значений ВОТ является смысловой компонент ‘связь описываемой ситуации с моментом речи’. Он вырастает из указания на пространственную близость демонстрируемого объекта к говорящему, которое присутствует в значении указательного ВОТ. Вследствие стандартного метафорического переноса ‘пространственные

[4] В частности, лексема ВОН 4 является обиходной. Кроме того, она обычно употребляется только тогда, когда А1 выражено существительным или личным местоимением, причем в вопросительных предложениях ее использование затруднено.

характеристики’→ ‘временные характеристики’ идея пространственной близости к говорящему трансформируется в идею темпоральной близости описываемой ситуации к моменту речи.

Указанный компонент в той или иной форме присутствует по крайней мере в 5 значениях ВОТ.

В некоторых случаях указание на связь с моментом речи присутствует прямо в толковании лексемы. Это относится к ВОТ 3.1, 3.2 и 8.1.

ВОТ 3.1 ‘Говорящий сообщает, что событие А1 произошло непосредственно **перед моментом речи** или произойдет **сразу после момента речи**’. Ср.

(22) *Вот только что <сию минуту> я его видел, и уже опять он куда-то исчез.*

(23) *Как это может быть, что вот только что он говорил с Берлиозом, а через минуту — голова…* (М. Булгаков).

(24) *Вот сейчас, сейчас, — говорил Павел. — Сейчас будет готово, накормим тебя* (А. Волос).

(25) *Пребывание с нею в одних стенах вызывало тихое озлобление, потому что постоянно чудилось: вот сейчас грохнется тарелка на пол, посыпятся книги с полки* (А. Азольский).

Похожее значение имеет также наречие *вот-вот*, имеющее толкование ‘через очень маленький промежуток времени после момента речи’: *Он вот-вот вернется*.

ВОТ 3.2 ‘Говорящий подчеркивает, что событие А1, которого он долго ждал, произошло или начало иметь место **незадолго до момента речи**, причем само это событие или его результаты сохраняются **в момент речи**’. Ср.

(26) *Вот и кончились каникулы*.

(27) *Вот ты и попался <Вот и случилось то, чего я всегда боялся>*.

(28) *Вот и познакомились <Вот и договорились; Вот мы и свиделись>*.

(29) *Гриша, дорогой, здравствуй! Вот и двухтысячный год на дворе. Вспомнил, что тебе стукнуло 60, и решил черкнуть письмо* (Г. Горин).

(30) *Вот и подстерегла меня любовь* (В. Астафьев).

(31) *Умывание короткое: каждая больная слегка намыливала руки, споласкивала их, потом лицо. И всё. Вот и до меня дошла очередь* (И. Грекова).

ВОТ 8.1. ‘Говорящий выражает сильное чувство А1, вызываемое у него какой-л. ситуацией, которую он рассматривает **в момент речи**, или сильное

чувство, вызываемое у него объектом или ситуацией А1, которую он рассматривает **в момент речи,** ожидая, что адресат начнет испытывать то же самое чувство' [А1 обычно восхищение, радость, возмущение или досада]. Ср.

(32) *Вот чудак! < Вот глупец!; Вот вздор!; Вот новость!; Вот невезение!; Вот загадка!>*

(33) *Вот замучил <надоел, пристал>!*

(34) *Вот ужас! <Вот стыд-то!; Вот досада !>.*

(35) *Вот радость! — воскликнула Анна и всплеснула руками* (А. И. Куприн).

(36) *Артем остановился* [...] *и подумал: «Вот чудо! Будто бы и не хотел в село идти, а пришел»* (Г. Марков).

В других случаях указание на связь с моментом речи вводится опосредованно, например, через указание на актуальность сообщаемой информации или указание на непосредственное восприятие событий.

Первое характерно для лексемы ВОТ 3.3. Она толкуется следующим образом: 'Говорящий указывает на то, что сообщаемая им информация А1, обычно касающаяся действий его самого или кого-л. из его личной сферы, является актуальной'. Ср.

(37) *— А вы опять вместе? — Да вот, в подъезде столкнулись, — отвечала Елена Николаевна* (А. Геласимов).

(38) *— Вам чего, ребята? — спрашиваю. — Да вот, — говорят, — пришли проведать, скоро ли «опохмелочная» откроется?* (Г. Горин).

(39) *— Василий Семенович, какими судьбами?* [...] *— Да вот, решил, знаете, справиться о здоровье. Как себя чувствуете?* (П. Галицкий).

(40) *— Куда летом едете? — Вот хотим в Болгарию рвануть на машине.*

(41) *— Как ваш младший поживает? — Да вот болеет опять <Да вот женился недавно>.*

В силу указания на актуальность информации данная лексема чрезвычайно характерна для разговорной речи и встречается, главным образом, в диалогах (причем часто употребляется после частицы *да*). Как правило, она используется, когда говорящий отвечает на вопрос адресата, в том числе невысказанный.

В работе В. Ю. Апресян « О временных значениях дейксисных слов» [Апресян 2014] описывается похожее употребление наречия *тут*. Ср.

(42) *Я тут сижу работаю.*

(43) *Я тут была в Лондоне.*

В подобных фразах *тут* значит ‘в момент, близкий к моменту речи говорящего’; в некоторых из них *тут* допускает замену на *вот (Вот сижу работаю; Вот была в Лондоне*), хотя и с легким сдвигом значения.

Указание на непосредственное восприятие событий входит в значение лексемы ВОТ 2, которая употребляется в нарративной речи и толкуется следующим образом: ‘Говорящий описывает события А1, которые произошли в прошлом или произойдут в будущем, как если бы он непосредственно воспринимал А1, чтобы адресат представил себе А1’. Ср.

(44) *Вот выйду на пенсию, поселюсь в деревне, буду выращивать кур.*

(45) *Вот, бывало, приду после работы домой, там уже стол накрыт.*

(46) *Но вот впереди появились защитники замка в малиновых накидках поверх камзолов* (В. Быков).

(47) *И вот однажды на рассвете, когда костёр догорел, небо едва забрезжило,* […] *он услыхал совсем рядом голоса* (А. Варламов).

В данном случае компонент ‘непосредственное восприятие’ имеет метафорическое значение и служит средством придания рассказу большей изобразительности (ср. использование настоящего исторического в той же функции).

Указание на связь с моментом речи в семантике ВОТ может преобразовываться в указание на связь с конкретной ситуацией общения. В частности, описываемый компонент присутствует в значении лексемы ВОТ 6.2, которая используется в иллюстративной функции и толкуется следующим образом: ‘Говорящий подчеркивает, что он переходит от общего утверждения А2 к конкретным разъяснениям или примерам А1’. Ср.

(48) *— Я понятия не имею, что такое Ничто.* […] *— Сейчас объясню. Вот предположим, ты умер. Можешь себе это представить?* (О. Ефремова).

(49) *Половину этих бумаг давно пора выбросить. Вот зачем тебе эта старая тетрадь?*

(50) *Я знаю наизусть все стихи из этого сборника. Вот хочешь, я тебе скажу, какое стихотворение на десятой странице?*

(51) *Так вести себя некрасиво. Вот скажи, разве тебе было бы приятно, если бы тебя стали дразнить плаксой?*

Рассматриваемая лексема обычно употребляется, когда говорящий хочет проиллюстрировать или разъяснить какие-л. утверждения конкретным

примером. Чаще всего с этой целью он предлагает собеседнику ответить на вопрос А1 (*Вот скажи…* — 150 вхождений в НКРЯ) или совершить действие А1, чаще всего — ментальное (*Вот допустим…* — 83 вхождения в НКРЯ; *Вот представь…* — 81 вхождение; *Вот возьмем…* — 35 вхождений; *Вот предположим* — 12 вхождений). Кроме того, *вот* может использоваться, когда говорящий предлагает совершить физическое действие, хотя это несколько менее типично (*Вот встань сюда <вот подойди сюда>*).

Таким образом, ВОТ 6.2 обычно указывает на переход от изложения абстрактных утверждений к общению с конкретным адресатом в рамках конкретной коммуникативной ситуации. При этом для ВОТ в описываемом значении чрезвычайно характерно употребление в сочетании с глаголами восприятия *смотреть, слушать, попробовать* в форме императива, обозначающими призыв к непосредственному восприятию информации:

(52) *Это очень вкусно. Вот попробуй.*

(53) *а) Я сейчас покажу тебе, как это делать. Вот смотри.*
б)«Этот барьер не прошли» — расшифруй, пожалуйста, — какой барьер? — Ну, вот смотри. Что такое шоу-бизнес? (А. Клейн).

Контексты такого типа очень многочисленны — сочетание *вот попробуй* зафиксировано в НКРЯ 50 раз, *Вот слушай* — 127 раз, *Вот послушай* — 109, *Вот смотри* — 291 раз. При этом контексты с сочетанием *Вот смотри* распадаются на 2 группы. В примерах типа *(53 а)* говорящий в буквальном смысле предлагает посмотреть на объект или ситуацию; в данном случае ВОТ 6.2 частично сближается с указательным ВОТ 1.1. (При этом при наличия паузы между *вот* и *смотри,* обозначаемой в тексте запятой, значение ВОТ может полностью сдвигаться в сторону ВОТ 1.1: *Вот, смотри, это Маша; Вот, смотри, бежит Вася).* В примерах типа *(53б)* о зрительном восприятии объекта речь не идет: в данном случае говорящий призывает адресата воспринять текст, который он намерен воспроизвести.

3.2. Временная близость и взаимообусловленность

Компонент ‘близость объекта к говорящему’ в значении ВОТ 1.1 может трансформироваться не только в компонент ‘близость к моменту речи’, о котором шла речь выше, но и в указание на временную близость двух событий, а также в указание на связь событий между собой. Этот смысл актуален, в частности, для лексемы ВОТ 4, которая толкуется следующим образом: ‘Говорящий подчеркивает, что желательное для адресата или говорящего событие А2 произойдет сразу после события А1, причем А1 невозможно без А2’. Ср.

(54) *Вот сделаю уроки и пойду гулять.*

(55) *Сейчас иду, вот только книжки уберу.*

(56) *Вот потеплеет, сразу же посадим картошку.*

(57) *Но всё-таки нехорошо, что я так распустилась, вот он уедет, и я покрашу* [волосы] (Л. Улицкая).

(58) *И всю дорогу приговаривала: — Вот приду, а папа дома! Просто ни о чем другом в тот день не могла говорить* (Д. Рубина).

(59) *Вот закончу последнюю работу и тогда обязательно займусь собой* (Э. Радзинский).

(60) *Вот скоро вырасту, тогда устрою радугу для всех — в полнеба!* (С. Георгиев).

Как мы видели, в значении ВОТ 4 смысловой компонент ‘временная близость событий’ соединен с компонентом ‘связь событий между собой’, причем второй, на наш взгляд, является производным от первого. В значении лексемы ВОТ 7[5] указание на взаимообусловленность событий представлено уже в качестве самостоятельного элемента, вне связи с указанием на их временную близость. Мы предлагаем следующее толкование ВОТ 7: ‘Говорящий подчеркивает, что А1 — следствие ситуации А2, о которой говорилось ранее, или вывод из А2’. Ср.

(61) *Он сломал велосипед, вот пусть теперь и чинит его.*

(62) *Погулял без шапки — и вот, пожалуйста, заболело горло.*

(63) *Не знаешь — вот и молчи.*

(64) *Вот и молодец <вот и хорошо>.*

(65) *Соседи скажут, что мы плохо работаем, вот и отнимут* [участок] (А. Варламов).

(66) *Их* [ульи] *у него украли, и он очень переживал, а сердце-то больное, вот и не выдержало* (Э. Лимонов).

(67) *Напился, как свинья, вот и не помнишь ничего!* (Я. Кудлак).

3.3. Правильность

Идея пространственной близости в значении ВОТ 1.1 может модифицироваться в идею содержательной близости, соответствия чему-л. — в частности,

---

[5] Данная лексема употребляется, главным образом, в сочетании с *и*, чаще всего — в постпозиции (*вот и*), реже — в препозиции (*и вот*).

близости чьих-л. слов или действий к истине, норме, правилу, — и, в конечном счете, в идею правильности.

Описываемый смысловой компонент мы находим в значении лексемы ВОТ 8.4[6], которая имеет следующее толкование: ‘Говорящий подчеркивает, что он считает слова или действия адресата правильными’.

(68) *Вот, правильно <Вот-вот, точно>.*

(69) *— Вот, вот, — кивал он;*

(70) *— Вот, правильно мать говорит, — сказала женщина-врач* (В. Голованов).

(71) *— Отдайте мое сочинение, Светлана Михайловна. — Вот правильно! Возьми и порви, я тебе разрешаю* (Г. Полонский).

(72) *Мэя сначала опустила руки, потом, виновато улыбаясь, подошла к нему. — Вот правильно! — он взял ее за руку, усадил в постель* (С. Осипов).

4. Частица *вон*. Метафоризация идеи удаленности

Для переносных значений ВОН оказывается актуален совершенно другой набор смыслов, не соотносимых с теми, о которых мы говорили применительно к ВОТ. При этом, как и в случае с ВОТ, эти смыслы возникают в результате модификации основного, указательного значения, точнее, компонента «удаленность от говорящего».

Этот компонент может переосмысляться в двух направлениях.

4.1. Высокая степень признака

Прежде всего, указание на большое расстояние, отделяющее объект от говорящего, может трансформироваться в смысловой компонент «высокая степень признака». Этот компонент является центральным для лексемы ВОН 1.2, которая имеет следующее толкование: ‘Говорящий обращает внимание адресата на высокую степень свойства или ситуации А1 или большое значение параметра А1, которое можно наблюдать в момент речи или которое адресат мог наблюдать незадолго до момента речи’. Ср.

(73) *Вон какой он ловкий — всех обвел вокруг пальца.*

(74) *Вон сколько с ним забот.*

---

[6] То же значение реализуется в сочетаниях *вот именно, вот то-то же.*

(75) *А у нас* [Томара] *будет сыта, обута, одета. За Таней вон сколько всего остаётся* (Л. Улицкая).

(76) *Ничего, растут. Дмитрия Борисыча вон как любят. Папой зовут, только я против* (И. Грекова). *Встаёт, задаёт ехидные вопросы, класс гогочет, а он сияет, вон, мол, какой я умник!* (Ю. Домбровский).

Кроме того, указание на высокую степень признака присутствует у ВОН в значении лексемы ВОН 2.2 (см. ее анализ ниже), в усилительном режиме употреблений, который обычно реализуется в сочетании со словами *где, куда, когда.* Ср.

(77) *Тот, другой Милюков, вон куда заехал — в Париж, а наш старался держаться поближе к дому* (Ф. Кривин) [говорящий считает, что герой заехал очень далеко].

(78) *Последствия абортов вон когда сказались,* [...] *когда располосовали её* (В. Астафьев) [говорящий считает, что последствия абортов сказались через большой промежуток времени].

(79) *Опять социальное неравенство. Ты у нас вон кто — официантка. А я-то всего-навсего — шнырь!* (Э. Рязанов) [говорящий считает статус официантки высоким].

4.2. Отсылка

Помимо идеи усиления, смысл «удаленность объекта от говорящего» порождает идею отсылки, которая фиксируется почти во всех производных значениях ВОН.

В большинстве случаев ВОН содержит отсылку к ситуации, которая известна говорящему. Этот компонент мы встречаем в значениях ВОН 2.2 и ВОН 1.2.

Начнем с лексемы ВОН 2.2. Отсылка к известной говорящему ситуации в этом случае является центральным компонентом значения; ср. толкование этой лексемы: ‘Говорящий подчеркивает, что адресат, как и сам говорящий, знает ответ А2 на вопрос А1, причем говорящий обычно не одобряет ситуацию А2’[7]. Ср.

(80) *Я с тобой по-хорошему, а ты вон что делаешь — школу прогуливаешь, родителям врешь.*

(81) *В Париже вон что творится! Коммуна!* (Л. Юзефович).

---

[7] Компонент ‘отсылка к информации, известной говорящему и адресату’ присутствует также в значении лексемы ВОН 4, указывающей на выделение объекта из множества однородных (о ней шла речь в п. 2.2). Наличие этого компонента делает нежелательным для ВОН 4 употребление в вопросительных контекстах. Ср. нежелательность замены ВОТ на ВОН в следующей фразе: *Вот вы, к примеру, что предпочитаете на обед?*

Из-за того, что ВОН 2.2 содержит отсылку к информации, известной адресату, указание на ситуацию А2 оказывается избыточным и обычно отсутствует. Ср.

(82) *У него денег, поди, побольше, чем у Суханова. Наверное, уж не один миллион в швейцарских банках. Вон что в газете про него пишут* (А. Белозеров).

(83) *Я ему помогал киль прилаживать, крылья клепал, красную звезду рисовал… А он вон что надумал* (А. Проханов).

(84) *А то ведь это прямо невозможно. У мамы приступ, а они вон что вытворяют* (В. Войнович).

(85) *Он — категорически против: только-только жить начали по-людски, а ты вон что затеяла* (И. Грекова).

Помимо ВОН 2.2 идея отсылки актуальна также для лексемы ВОН 3, выполняющей иллюстративную функцию. Ее толкование выглядит следующим образом: ‘Говорящий ссылается на факт А1, приводя его в качестве доказательства или иллюстрации утверждения А2, которое было сделано ранее’. Отличие этого значения от предыдущего, содержащего аналогичный элемент, состоит в том, что в данном случае ситуация, к которой говорящий отсылает адресата, может быть новой для последнего. Ср.

(86) *Совсем дети распоясались. Вон, мой младший вчера бабушку до слез довел.*

(87) *Без хозяина — вовсе конец. Поставят абы кого… Вон в Грачах. Поставили бабу — и за ночь разнесли мастерскую. Всё дочиста* (Б. Екимов).

(88) *Скажи спасибо, что хоть исправить дают заранее. Бывает хуже. Вон, Муза рассказывает…* (А. Солженицын).

Данная лексема частично соотносится с лексемой ВОТ 6.2, которая уже обсуждалась выше (см. 3.1). ВОТ 6.2 и ВОН 3 сближает указание на переход от утверждения к его иллюстрации; в некоторых контекстах рассматриваемые лексемы даже являются взаимозаменяемыми. Ср. следующие фразы, в которых ВОН легко может быть заменено на ВОТ:

(89) *Без хозяина — вовсе конец. Поставят абы кого… Вон в Грачах. Поставили бабу — и за ночь разнесли мастерскую.*

(90) *А чего это ты меня разжалелся? Что я — дефектная! Вон у меня подруга — так у нее ноги кривые, будто она на цистерне до Киева ехала — вот ее и жалей* (Э. Радзинский).

(91) *Всякое отступление от линии партии — это смерть или предательство. Вон, какие люди были, а как скатились в болото оппозиции,* [...] *то вон к чему и пришли!* (Ю. Домбровский).

Различие между ВОТ и ВОН в иллюстративном значении связано с тем, что для ВОН обязательной является идея отсылки к какому-л. факту, который говорящий использует в качестве иллюстрации своего утверждения. Между тем, ВОТ может употребляться, когда говорящий с той же самой целью (для разъяснения и иллюстрации только что сделанного утверждения) задает вопрос адресату, предлагает совершить какое-л. действие и пр. Ср. следующие фразы, в которых замена ВОТ на ВОН была бы невозможна:

(92) *Так вести себя некрасиво. Вот скажи, разве тебе было бы приятно, если бы тебя стали дразнить плаксой?*

(93) *Вот представь… <Вот попробуй…>*

(94) *Я понятия не имею, что такое Ничто.* […] *— Сейчас объясню. Вот предположим, ты умер. Можешь себе это представить?*

Невозможность употребления ВОН 3 в ситуации, когда говорящий для иллюстрации своей мысли предлагает совершить какое-л. действие, полностью подтверждается корпусными данными. Так, сочетания *Вон допустим.., Вон представь.., Вон предположим.., Вон попробуй..* не встретились в корпусе ни разу, сочетание *Вон скажи..* появилось всего 2 раза (при том, что *Вот скажи* зафиксировано в 150 случаях). Сочетание *Вон слушай..* встретилось 1 раз, причем в значении, близком скорее к ВОН 1.1, чем к ВОН 3: *А ты вон слушай — музыка, — позёвывая, советует ему рябой, коренастый товарищ* (М. Горький). В данном случае перед нами ослабленное употребление ВОН 1.1 в значении ‘Говорящий привлекает внимание адресата к звукам А1’. Здесь ВОН, в отличие от ВОН 3, выступает не в функции иллюстрации, а в функции демонстрации.

5. Выводы

Анализ многозначности ВОТ и ВОН демонстрирует значительное расхождение в метафорических употреблениях, обусловленное различием в дейктической семантике мотивирующих значений ВОТ 1.1 и ВОН 1.1.

При этом компонент ‘близость объекта к говорящему’, входящий в значение ВОТ 1.1, чаще всего метафорически переосмысляется в компонент ‘близость к моменту речи’ (его вариантами являются компоненты ‘актуальность’, ‘непосредственное восприятие событий’, ‘связь с говорящим и ситуацией общения’). В данном случае перед нами — вполне типичное для языка преобразование пространственного дейксиса во временной.

В свою очередь, компонент ‘близость к моменту речи’ в значениях ВОТ проходит в процессе своей трансформации следующие этапы, каждый из который характеризуется нарастанием метафоричности: ‘близость к моменту речи’ (временной дейксис) → ‘временная близость событий’ (таксисное значение) → ‘взаимообусловленность событий’.

Что касается частицы ВОН, то смысл ‘удаленность объекта от говорящего’, входящий в значение ВОН 1.1, трансформируется в смысл ‘высокая степень признака’, а также в указание на отсылку к какому-л. факту (обычно известному говорящему).

Литература

1. *Борисова, Овчинникова 2011* — Е. Г. Борисова, Т. Е. Овчинникова. Параметр близости в метафорическом пространстве // Компьютерная лингвистика и интеллектуальные технологии / По материалам ежегодной международной конференции «ДИАЛОГ» (2011). Вып. 10. М., 2011. С. 153–158.
2. *Апресян 2014* — В. Ю. Апресян. Тут, здесь и сейчас. О временных значениях пространственных дейктических слов. // Русский язык в научном освещении. М., 2014, 27. С. 9–41.

References

1. *Borisova, Ovcyinnikova 2011* — E. G. Borisova, T. E. Ovcyinnikova. Parameter of nearness in a metaphorical space [Parametr blizosti v metaforicheskom prostranstve] // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog 2011” [Komp’yuternaya Lingvistika i Intellektual’nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii “Dialog 2011”], Vol. 10, Bekasovo, pp. 153–158.
2. *Apresyan 2014* — V. U. Apresyan. «Тут», «здесь» and «сейчас». About the temporal values of the spatial deictic words. [Tut, zdes’ i seychas. O vremennykh znacheniyakh prostranstvennykh deikticheskikh slov] // Russian language in scientific lighting [Russkiy yazyk v nauchnom osveshchenii]. M., 2014, 27, pp. 9–41.

АВТОМАТИЧЕСКОЕ ПОПОЛНЕНИЕ БАЗЫ ИМЕНОВАННЫХ СУЩНОСТЕЙ НА ОСНОВЕ ПОЛЬЗОВАТЕЛЬСКИХ ЗАПРОСОВ

Кудинов М. (m.kudinov@samsung.com),
Пионтковская И. (p.irina@samsung.com)

Исследовательский Центр Самсунг, Москва, Россия

В работе описан процесс пополнения базы именованных сущностей, предназначенной для поддержки голосовых команд на устройстве. Пополнение производится полностью автоматически без участия редактора путем анализа логов запросов к диалоговой системе. Логи состояли из откликов системы распознавания речи и поэтому содержали большое количество ошибочных распознаваний именованных сущностей. Поиск подобных ошибок также осуществлялся в рамках описываемого подхода. Используются методы на основе взаимной информации и скрытых марковских моделей.

Ключевые слова: скрытые марковские модели, редакторское расстояние, система распознавания речи, именованные сущности, анализ поисковых запросов

AUTOMATIC UPDATE OF THE NAMED ENTITIES DATABASE BASED ON THE USERS QUERIES

Kudinov M. (m.kudinov@samsung.com),
Piontkovskaya I. (p.irina@samsung.com)

Samsung R&D Institute, Moscow, Russia

We describe an algorithm of update of the database of named entities providing support of voice commands on a device. The update is made automatically with no human assistance by means of analysis of query logs of the dialogue system. The logs consisted of responses of the automatic speech recognition engine and thus contained erroneous recognitions. The search of such mistakes is also made as a part of our method. The problem of named entities extraction was solved by means of an algorithm based on entropy and mutual information statistics. The detection of recognition mistakes was made by means of a novel data-driven probabilistic approach taking into

account grapheme substitution statistics in the data. Assuming grapheme alignment hidden, we use the EM algorithm for training the model. As a result we obtain a statistical model capable for sequence similarity assessment. The algorithm based on our similarity score performs better in terms of F_1-measure than one using the classical Levenshtein distance.

Keywords: Hidden Markov models, edit distance, speech recognition, named entities, search query analysis

1. Problem statement. Dialog system

Consider the following simple scheme of the voice commands engine working as follows:

1) The user's query is recorded using the embedded microphone. The recorded speech signal is input to the speech recognition system which outputs text string sometimes containing errors.
2) The text is input to the query grammar based parser.
3) The parser outputs the semantic frame which determines the system response.

The system needs named entities database for parsing of queries like

покажи мне мультфильм пингвиненок пороро / show me the cartoon pororo the little penguin

я хочу посмотреть новый сезон сериала ментовские войны / I want to watch the new season of the series cops wars

The problem of the database support may be solved by means of collecting a gazetteer grabbing named entities from program guide or film ads but this approach has two complementary drawbacks: a) such list would have contained shows no one ever watches; b) the list would have missed popular video clips on YouTube, Vimeo etc.

The other problem caused by gazetteer based approach and an external speech recognition engine (ASR) is the need of interpreting queries containing ASR errors. A trivial analysis has shown that for a series Magnificent Age popular among Russian housewives the proportion of queries where the film name undergoes different changes reaches 5%. Adding these variants to the gazetteer is a necessary measure for providing good performance of the service.

Thus, we have to solve two tasks:

1) Find film names (TV programs, actors and directors names etc.) absent in the gazetteer and spelled right;
2) Find different variants of recognition of the objects in the gazetteer and add them with corresponding mark.

The problems will be considered below in corresponding sections.

2. Named entities extraction

The problem of the multi-word named entities extraction in short queries may be easily converted into the problem of collocation extraction. The classical approach to this task is the calculation of the pointwise mutual information between tokens:

$$PMI(x;y) = \log \frac{f(x,y)}{f(x)f(y)},$$

where $f(t)$ is the frequency of the token sequence t in corpus. PMI measures mutual dependence between two random events (x,y). All the token pairs found in the corpus were included into the list sorted by calculated PMI. It was the way the top list for the collocations of length 2 was made up. For the search of sequences of three tokens we take minimum of two calculated PMI scores:

$$PMI(x;y;z) = \min\big(PMI(x;yz), PMI(xy;z)\big).$$

The PMI for longer sequences was calculated the same way. We made experiments with the sequences of the length up to 6 (*Чип и Дейл спешат на помощь/ Chip ‹n Dale Rescue Rangers*). The top list on PMI for the sequences of length from 2 to 4 is shown in the column 1 of the table 1. It is obvious that some sequences are prefixes and suffixes of others.

The method of sequence merging was based on the notion that the $PMI(w_1 \dots w_t)$ of the named entity sequence is: a) greater than the $PMI(w_1 \dots w_{t-1})$ of its prefix; b) greater than the $PMI(w_1 \dots w_{t+1})$ of the sequence $w_1 \dots w_{t+1}$ with the prefix $w_1 \dots w_t$.

The experiment has shown that such notion leads to good results.

We used the following algorithm for n-gram merge. Firstly, the n-gram prefix tree was constructed such that every partial way corresponded to a n-gram and each node contained corresponding calculated PMI.Further, the tree was traversed such that each time the PMI of longer path was less than the PMI previous partial path the algorithm output the partial way and switched to the next branch. The sequences having a common suffix were merged according to the same principle.

Fig. 1. Prefix tree for n-gram merge

We also did the same operation in the reverse order with the suffix tree. We had to do it because the forward pass missed the sequences with high entropy of the prefix: **Three** days to kill.

The result of the described algorithm is shown in the column 2 of the table 1.

The last step we need to extract named entities is separate true named entities from other frequent collocations. There are at least two approaches to this problem. The first one is to take the hypothesis that named entities have similar distributions of the neighboring words in the query. For example, the cartoons and TV shows are often found with the words *watch, season, episode* etc. It would make possible using of any reasonable linear classifier e.g. logistic regression or SVM. But we revealed that the named entities may be detected reasonably well only according on the entropy of this distribution. As far as entropy characterizes the degree of "uniformness" of the distribution, the distributions with high peaks should have lower entropies. It is obvious that the word *film* which potentially can be found in context of every named entity with equal probability, should have greater entropy than a series name. After the final sort on entropy the top list looked like in the column 3 of the table 1.

It is easily seen that the top list is rather homogenous sample of film names, actors names and pop stars. At the same time we see that the most popular films did not get to the top list because they had too many different contexts. This result has an important advantage because it detects named entities which have little chance to get to a manually collected gazetteer. The other remarkable issue here is the top-1 entry which is not named entity. This strange result is caused by abnormal frequency of the n-gram *чужие два фантастика боевик*: this whole request was encountered 15 times of all 20 appearances of the film *Aliens* in the search log.

Table 1. N-gram top list

Sort by PMI	**After n-gram merge**	**Sort by Entropy**
маша и медведь	маша и медведь	чужие два фантастика боевик
маша и	http //www google com	серая шейка
и медведь	www yandex ru	спортлото восемьдесят два
http //www google com	три d	самолеты огонь и вода
www yandex ru	две тысячи четырнадцать	руби и йо йо
* *	* *	феи загадка пиратского острова
три d	улыбка пересмешника	секс по дружбе
две тысячи четырнадцать	vk com	мистер и миссис смит
http //www google	новые серии	фильмы для взрослых
//www google com	все серии	teenie лав
yandex ru	физрук второй сезон	огги и тараканы
www yandex	www яндекс ru главная	полли робокар

Sort by PMI	After n-gram merge	Sort by Entropy
vk com	экстрасенсов пятнадцатый сезон	мультфильм тини лав
улыбка пересмешника	смотреть онлайн	крошка енот
все серии	битва экстрасенсов	отпуск по обмену
смотреть онлайн	google com	осторожно обезьянки
второй сезон	чернобыль зона отчуждения	бен и холли
физрук второй сезон	сын за отца	тотальная распродажа
тысячи четырнадцать	д * *	блондинка в эфире
новые серии	ха ха ха ха ха	идентификация борна
все серии подряд	ну погоди	дьявол носит prada
www яндекс ru главная	свинка пеппа	люди икс дни минувшего будущего
http //www	серии подряд	федорино горе
google com	человек паук	большое зло и мелкие пакости
//www google	ж * * у	новогодние приключения маши и вити
б * *	в ж * * у	пингвинёнок пороро
битва экстрасенсов	финес и ферб	паровозик томас и друзья
* * а	черепашки ниндзя	по дороге с облаками
д * *	физрук два сезон	котёнок по имени гав
маша и медведь	маша и медведь	притворись моей женой

3. ASR system errors detection

Let us now turn to the second problem. Assume that we managed to extract new good named entities and we have log of the ASR engine. Assume now that we can find word sequences according to the mask *I want to watch the film <FILM NAME>*. We believe that the placeholder FILM NAME is filled with an actual name of some movie. Now if the FILM NAME is not present in the gazetteer and its frequency is low we consider it as a candidate to the list of wrong recognitions of named entities and we must match it against gazetteer entries.

The last problem is almost classical statement of the sequence alignment problem which often emerges in the natural language processing [6] and speech recognition [4]. Speech recognition is also connected with grapheme-to-phoneme conversion ([2], [3], [1]), another important sequence alignment problem. [1] describes the approach based on so called graphones i.e. letter-phoneme pairs with and Baum-Welsh-like learning algorithm.

This approach leads to the method of similarity distance measure alternative to the popular Levensthein distance. Levenshtein distance gives equal penalty $d(x,y)$ to any deletions, insertions or replacements in the string S_1 relative to S_2. But the errors made by ASR systems tend to emerge because of replacement of a word w_1 with similarly

sounding word w_2. In that sense it would be reasonable to use a metric where $d(x,y)$ is higher for those pairs of letters x,y, which are more likely to be in replacement pairs ("o"–"a", "м"–"н" etc.), and lower otherwise. It is also logical to take the data driven approach where $d(x,y)$ is calculated from data. Then we can take the probability of replacement of the letter x with the letter y in the recognized string as a measure $d(x,y)$. We will test the performance of the Levenshtein measure with the one proposed here.

Consider the probability that the string S will be recognized as a string R:

$$\mathbf{P}(S,R) = \sum_{\{L_{S:R}\}} \mathbf{P}(S,R,L)$$

where $\{L_{S:R}\}$ is a set of all alignments between S and R.

Let $L_{S:R} = l_1, l_2, \dots l_n$ be the sequence of replacements of substrings in S with the substrings in R. Then

Alignment $L_{S:R}$ is a latent sequence like allophone models sequence in HMM-based speech recognition. Taking reasonable upper limit on the length of l_i it is possible to use a modified version of EM-algorithm, similar to *Baum-Welch* [5] as proposed in [1]. Moreover, in place of $\mathbf{P}(l_i)$ a bigram probability $\mathbf{P}(l_i \mid l_{i-1})$ may be taken, thus forming the model of higher order.

We calculate Levenshtein similarity measure as normalized Levenshtein distance LD between candidate and pattern strings:

$$L_1 = \frac{LD}{N}$$

where N is the length of the candidate. To calculate the probability based similarity measure we also should reckon difference of lengths and the fact that $\mathbf{P}(S,S) \neq 0$, which is not true for the valid distance:

$$L_2 = \left| \frac{\log\big(\mathbf{P}(S,S)\big)}{M} - \frac{\log\big(\mathbf{P}(S,R)\big)}{N} \right| + |M - N| \cdot \alpha,$$

where M and N are accordingly lengths of the pattern and the candidate, α is the penalty for difference between string lengths used as a hyperparameter.

The criterion of taking the decision that R is wrong recognition of S was taken if the measures L1 and L2 exceeded corresponding thresholds.

4. Experiments

In our experiments we used the dialog system query log. We chose the queries which system could not respond and extracted those which contained named entities.

The initial sample comtained 15,000 different queries. We took the queries with corpus counts above 20 and excluded TV control queries (poweroff, next channel etc.). Based on the popular queries we automatically obtained regular expressions for the extraction of the candidate strings.

There were 916 pairs named entity-recognized string in the training set and 89 pairs in the test set.

Both algorithms took as an input the list of candidate strings. The output was the list of pairs named entity-recognized string. Based on the lists in the test set and the algorithm outputs we calculated precision, recall and F_1-measure.

The results of the experiments for Levenshein-based algorithm and two versions of the HMM-based algorithm with first and second oreder probabilities are given below:

Method	Precision	Recall	F_1-measure
Levenshtein distance	0.76	0.72	0.74
Unigrams	0.85	0.98	0.91
Bigrams	0.83	0.97	0.89

Poor performance of the bigram model may be the result of overfitting.

5. Discussion

We proposed the algorithm for the gazetteer update based on the users' queries to the dialog engine. The proposed solution is able not only update a gazetteer but also detect different variants of wrong recognitions returned by the external ASR system.

The HMM based similarity measure allows to get more accurate predictions when we try to match some candidate string against patterns in the gazetteer. The key advantage of the method is that due to the use of the probabilistic measures of string transformation the similarities between strings transformed by means of likely replacements are less than for strings transformed by means of rare replacements. The algorithm for HMM-based similarity calculation performs better in terms of F1-measure than widely used Levenshtein distance.

References

[1] *Bisani M., Ney H.,* (2008), Joint-sequence models for grapheme-to-phoneme conversion, Speech Communication, 50 (5), 434–451.

[2] *Lucassen J. M., Mercer R. L.,* (1984) An information theoretic approach to the automatic determination of phonemic baseforms, Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP'84, Vol. 9, pp. 304–307

[3] *Luk R. W. P., Damper R. I.,* (1996), Stochastic phonographic transduction for English, Computer Speech & Language, 10(2), 133–153

[4] *Rabiner L., Rosenberg A., Levinson S.,* (1978), Considerations in Dynamic Time Warping Algorithms for Discrete Word Recognition, The Journal of the Acoustical Society of America, 63(S1), S. 79-S79.

[5] *Rabiner L.,* (1989), A tutorial on hidden Markov models and selected applications in speech recognition, Proceedings of the IEEE, 77(2), 257–286.

[6] *Schulz Klaus U.; Mihov Stoyan,* (2002), Fast string correction with Levenshtein automata, International Journal on Document Analysis and Recognition, 5(1), 67–85.

ВАЛЕНТНОСТИ АБСТРАКТНЫХ СУЩЕСТВИТЕЛЬНЫХ: РЕДУКЦИЯ VS. КОНКРЕТИЗАЦИЯ[1]

Кустова Г. И. (galinak03@gmail.com)

Институт русского языка им. В. В. Виноградова РАН, Москва, Россия;
Национальный исследовательский Томский государственный университет, Томск, Россия

Абстрактные существительные разных семантических классов (интерпретация: *неприятность*, *промах*, *разлад*; мероприятие: *референдум*, *совещание*, *турнир*; деятельность: *меры / работа / усилия* [*по урегулированию*] и др.) рассматриваются по аналогии с речевой и ментальной лексикой, которая имеет валентность содержания, выражаемую обычно придаточным (*думал, что P*; *рассказал, что P*), и валентность темы, выражаемую предложной группой (*думать / мысли о чем*; *рассказать / рассказ о чем*).

Исследуется валентность темы у разных классов абстрактных существительных и особенности ее выражения предложной группой *по* + Дат. п. (*механизм по привлечению новых клиентов*; *негативная тенденция по ухудшению портфеля*; *шаги по осуществлению своего мандата*).

Тема, с одной стороны, является редукцией содержания, но в то же время это и конкретизация, поскольку без такой денотативной привязки сообщение с абстрактной лексикой становится неинформативным.

Ключевые слова: абстрактная лексика, предложная группа *по* + Дат. п., валентность содержания, валентность темы, ментальная и речевая семантика

ABSTRACT LEXEMES' VALENCIES: REDUCTION VS. SPECIFICATION

Kustova G. I. (galinak03@gmail.com)

Vinogradov Russian Language Institute of the Russian Academy of Sciences, Moscow, Russian Federation;
National Research Tomsk State University, Tomsk, Russian Federation

[1] Работа выполнена при поддержке РГНФ, проект № 14-04-00507а.

Abstract vocabulary of different semantic classes (interpretation: *nepriyatnost'* ('nuisance', 'trouble', 'annoyance'), *promakh* ('a piece of carelessness'), *razlad* ('discord'); event: *referendum*, *soveshchanie* ('meeting'), *turnir* ('tournament'); activities: *mery* ('measures') / *rabota* ('work') / *usiliya* ('efforts') [*po uregulirovaniyu* ('on the settlement')] et al. are considered by analogy with speech and mental lexemes. The latter lexemes have valency on content (it is usually expressed by the subordinate clause) and valency on topic (it is usually expressed by the prepositional phrase *o X* 'about X').

Abstract lexemes valency on content / topic may also be expressed by the prepositional phrase *po X* [Dat.] 'on X': *mekhanizm po privlecheniyu klientov* 'mechanism to attract customers', *negativnaya tendentsiya po ukhudsheniyu portfel'a* ('negative trend for the deterioration of the portfolio'), *shagi po osushchestvleniyu mandata* ('steps to implement the mandate').

Valency on topic is both a reduction and a specification of the content of the situation.

Key words: abstract vocabulary, prepositional phrase, valency on content, valency on topic, speech and mental semantics

Валентности абстрактной и конкретной лексики

Абстрактная лексика имеет существенно другие свойства, чем лексика конкретная. Это касается не только лексических значений, но и семантического наполнения грамматических категорий. Например, падеж, хотя он в целом является синтаксической категорией, тем не менее имеет некоторую семантику, причем наиболее отчетливо семантическое содержание падежей можно сформулировать для конкретной лексики.

Семантика падежных форм при конкретных (физических) предикатах коррелирует с семантическими ролями их актантов (хотя прямой корреляции между ролью и падежом в русском языке нет). Сама номенклатура основных ролей — агенс, пациенс, инструмент, место — ориентирована на физические ситуации и предметные актанты. Выбор падежной или (особенно) предложной формы является семантически мотивированным и соотносимым с денотативной ситуацией, см., например, [Wierzbicka 1980].

К абстрактной лексике эта номенклатура ролей применима в очень слабой степени. Во-первых, абстрактная лексика, по своей природе, является результатом «отвлечения» от денотативно-референтного уровня, от мира физических предметов и процессов. Во-вторых, предложения, в которых фигурирует абстрактная лексика, часто являются результатом синтаксического преобразования, усложнения или компрессии некоторых исходных структур. Например, творительный падеж в описаниях физических ситуаций часто имеет значение инструмента, ср. *зачерпнул воды* ***ковшом***. Однако в предложении *Он закончил свою речь* ***обращением*** *к зрителям* невозможно прямо приписать «денотативную» семантику творительному падежу, тем более что это предложение можно рассматривать как результат синтаксического преобразования некоторой исходной конструкции: *В конце своей речи он* ***обратился*** *к зрителям* — где смысл

‘обращение’ вообще не представлен падежной формой. Тем не менее и на уровне абстрактной лексики валентности могут иметь содержательное наполнение и, следовательно, существуют какие-то семантические закономерности и ограничения, регулирующие выбор управляемых падежных и предложных форм, поскольку в каждом случае мы употребляем не любую, а определенную форму.

В данной работе мы рассмотрим одну из форм выражения валентности абстрактных существительных. Речь пойдет об абстрактных существительных с предикатной семантикой. Это могут быть слова, которые образованы от предикатов — глаголов и прилагательных (ср. *поддержка*, *попытка*, *разбирательство*, *неизвестность*) — или просто имеют глагольный корень, не имея соотносительного глагола (ср. *сделка*, *обстановка*); это могут быть и заимствованные слова (*семинар*, *политика*, *эксперимент*, *проблема*), — важно, что они отсылают к ситуации.

В отличие от глаголов, для которых управление — это постоянная (словарная) характеристика, управление имен вообще является вторичным в том смысле, что заранее неизвестно, каким падежом или предложной группой должно управлять данное производное или заимствованное существительное. Конечно, существительное может управлять инфинитивом (ср. *решение уехать*; *цель достичь согласия*) или может наследовать управление от производящего предиката (ср. *надеяться на что — надежда на что*, *уверен в чем — уверенность в чем*). Но нас будут интересовать способы выражения валентностей падежными формами, специфичные для существительных.

По Е. Куриловичу [Курилович 1962], у существительных (как конкретных, так и абстрактных) есть универсальный синтаксический падеж — родительный, за которым могут скрываться самые разные семантические отношения: субъекта (*дежурство врача*), объекта (*вызов врача*), обладателя (*платье Маши*), носителя признака (*прозрачность стекла*), автора (*портрет Репина*), образа (*портрет Толстого*) и др.

У Е. Куриловича речь идет, в основном, о предметных валентностях (т. е. соответствующих предметным актантам): *строительство* ***дома*** (*строить дом*), *приезд* ***писателя*** (*писатель приехал*). Впрочем, родительным могут выражаться и пропозициональные (и вообще непредметные) валентности абстрактных существительных, ср.: *угроза* ***отставания***, *профилактика* ***нарушений***. Однако такой родительный семантически опустошен.

> **Примечание.** Кроме того, непредметная валентность может выражаться творительным беспредложным, ср. *управление* ***процессами***, *недовольство* ***результатами***. Этот материал в данной работе рассматриваться не будет.

Нас будут интересовать более семантически нагруженные способы выражения валентностей абстрактных существительных с помощью предложных групп, которые имеют пусть «слабую», но все-таки различимую семантику.

«Денотативная привязка» абстрактных существительных и способы ее выражения

Наиболее востребованными в качестве средств выражения валентностей в сфере абстрактной лексики являются предложные обороты (далее для краткости будем говорить просто «обороты») *по* + Дат[ельный] и *в* + Пр[едложный], более ограниченное употребление у оборота *с* + Твор[ительный]. В каких-то случаях они конкурируют между собой: *отставание в отдельных видах продукции — по отдельным видам продукции — с отдельными видами продукции*; *неприятности в бизнесе — с бизнесом — по бизнесу*; в каких-то случаях — с другими формами: *программа по переустройству — программа переустройства*; *информация по встрече в верхах — информация о встрече в верхах*; *решение по продаже акций — решение о продаже акций — решение продать акции*.

Примечание. Разумеется, эти обороты употребляются и с глаголами (ср. *превосходить / опережать в чем / по чему / с чем*), при этом они не являются «исконными» валентностями глаголов, а появляются у производных (абстрактных) значений, см. [Кустова, в печати].

Семантика этих оборотов определяется, во-первых, семантикой предлога, а во-вторых, тем, что эти формы по-разному взаимодействуют с семантическими классами имен: одни существительные предпочитают *в* + Пр (*верность*, *диссонанс*, *дисгармония*, *напряжение*, *передышка*, *оплошность* ***в чем***), другие — *по* + Дат (*прогнозы* ***по поставкам*** *X-а*; *операция* ***по поиску*** *X-а*; *усилия* ***по обнаружению*** *X-а*), наконец, третьи (*неудача*, *осложнение*, *перевес*, *несогласие*, *разлад*, *успехи*) сочетаются с обеими формами (*несогласие* ***по вопросу*** */* ***в вопросе*** *X-а*). Формы *в* + Пр и *по* + Дат в контексте абстрактных существительных обозначают «денотативную привязку». Говоря о денотативной привязке, мы имеем в виду следующее.

Абстрактные существительные являются осмыслением, интерпретацией (в широком смысле) некоторой конкретной денотативной ситуации. Это очевидно для оценочных слов типа *злоупотребления*, *нарушения*, *затруднения*, *помехи*, *неприятности*, *притеснения*, *неравноправие*, *произвол*, *польза*, *помощь*, *заслуги* и под. Все они предполагают некоторую денотативную расшифровку: в чем именно состоят неприятности, нарушения, заслуги и т.д. Однако и нейтральные слова типа *усилия*, *попытки*, *шаги*, *меры*, *инициатива* тоже отсылают к ситуации, которая происходит (или будет происходить) во времени и имеет конкретное денотативное содержание. Часто бывает необходимо расшифровать это денотативное содержание или дать хотя бы краткое указание на то, к какой сфере относится эта конкретная ситуация.

Приведенный выше набор падежных и предложных форм, начиная с полностью синтаксического родительного и заканчивая более семантизированными предложными оборотами, как раз и используется в функции выражения такой «денотативной привязки».

Примечание. Разумеется, у этих форм есть и другие функции, однако они сейчас не обсуждаются.

Валентности содержания и темы у слов с ментальной и речевой семантикой и других абстрактных слов

Типы денотативной привязки могут быть различными, в силу чего различна и семантика форм, ее выражающих. У некоторых абстрактных слов денотативная привязка осуществляется через валентность сферы (о термине «валентность сферы» см. [Апресян и др. 2010: 376]; эту валентность, выражаемую формой *в* + Пр, мы рассматриваем в другой работе, см. [Кустова, в печати]). Сейчас нас будут интересовать слова, имеющие валентность содержания и / или темы.

Сначала обратимся к тем классам слов, для которых валентность содержания и темы является исконной, — это предикаты со значением речи и мысли (как глаголы, так и существительные).

Предикаты этих классов (которые являются разновидностью абстрактных предикатов), как уже говорилось, отличаются от «обычных» (физических) предикатов по целому ряду параметров. Так, два «обычных» предиката — это два отдельных сообщения. В предложениях с абстрактными предикатами, например ментальными[2] или речевыми, тоже два предиката — один в главной клаузе, другой — в пропозициональном объекте (*Он* ***подумал****, что гости не* ***придут***; *Он* ***сказал****, что гости не* ***придут***). Однако предикат мысли или речи — это не отдельное сообщение, а модус, рамка для пропозиционального объекта. «Думать» — это и значит иметь какие-то мысли. Т.е. пропозициональный объект (содержание) — так сказать, конкретизация, расшифровка «думать». «Говорить» — это и значит произносить то предложение, содержание которого изложено в придаточном. Именно поэтому придаточное, выражающее пропозициональный объект, называется валентностью содержания. Очевидно, что у предикатов физических действий отношение с объектом совсем другое: в структуре «*резать хлеб*» актант *хлеб* не является содержанием предиката *резать*.

Наряду с валентностью содержания речевые и ментальные предикаты имеют валентность темы, «исконным» выражением которой является группа *о* + Пр: *отчет / доклад* [*доложить*] ***о работе***; *мысли* [*думать*] */ идеи* ***о переустройстве***.

Тема — это, вообще говоря, тоже передача содержания, но очень редуцированного — в виде самого общего представления. Т.е. тему можно считать редукцией исходного «полного» содержания (пропозиционального объекта, выраженного придаточным).

Абстрактные слова некоторых семантических классов можно рассматривать по аналогии с ментальными и речевыми. Такие слова, как уже говорилось выше, относятся к какой-то конкретной денотативной ситуации (которую они осмысляют и интерпретируют с определенной точки зрения и которую можно считать их денотативным (пропозициональным) содержанием) и предполагают

---

[2] Термины «ментальные предикаты», «ментальная лексика» и под. используются в статье в традиционном лингвистическом значении, а не в смысле психолингвистического «ментального лексикона».

расшифровку или хотя бы условное обозначение денотативного содержания, отсылку к нему: *Для того чтобы избежать* ***скандалов по обвинениям*** *в применении допинга, представляется необходимым сделать следующее* [«Известия», 2003.02.07][3] — обвинения в применении допинга — это и есть скандальная ситуация, содержание скандала; *Допущены серьезные* ***просчеты по оптимизации*** *размещения по территории профессиональных колледжей* [РИА Новости, 2009.08.03] — недостаточная оптимизация — это и есть просчет. Для таких предикатных слов содержание и тема — практически одно и то же, — в том смысле что у них используется одна форма для передачи того и другого, поскольку содержание трудно выразить с помощью придаточного (разве что через посредство вспомогательного глагола: *Просчеты* ***состоят в том, что P***).

Распространенными синонимами *по* + Дат являются обороты с производными предлогами *по поводу X-а*; *по отношению к X-у* (иногда также — *по случаю*, *по адресу*, *насчет* и др.), ср.: *Но дискуссия по «Славнефти» не была предана огласке ни одной из сторон. В отличие от* ***скандала по ФСФО***. [«Еженедельный журнал», 2003.03.24] vs. — *Как вы относитесь к* ***скандалу по поводу вашего первого проекта?*** [«Известия», 2003.01.19] — здесь *проект* — именно тема, «заголовок» скандала, т. к. неизвестно, какие именно скандальные события связаны с проектом.

Такая абстрактная лексика часто вообще не имеет никакой исконной валентности (наподобие *о* + Пр) и вынуждена «выбирать» из имеющегося в языке арсенала форм и оборотов наиболее подходящие для выражения темы / содержания. Анализ языкового материала показал, что несколько достаточно представительных семантических групп абстрактной лексики для выражения темы используют оборот *по* + Дат.

Семантические классы абстрактных существительных с валентностью темы / содержания *по* + Дат

В качестве исходного материала был взят список абстрактных лексем, сформированный на основе словаря Ожегова-Шведовой. Под лексемой мы, вслед за Ю. Д. Апресяном, понимаем слово в отдельном значении. Так, слово *ракурс* имеет в этом списке 3 вхождения (3 лексемы, т. е. 3 разных значения), слово *положение* — 5 вхождений (лексем). Всего в списке абстрактных лексем — около 10 тыс. единиц. Из них путем сплошной выборки были отобраны лексемы, у которых управление *по* + Дат подтверждается материалами Национального корпуса русского языка (НКРЯ). Их оказалось около 700, т. е. порядка 7 % (речь идет именно об управлении, т. е. выражении валентности, а не об адвербиалах типа *по принципу*, *по плану*, *по контракту*, *по методу*, *по графику*, *по решению*, *по инициативе* и т. п.).

Затем из этих 700 были исключены группы параметрических слов: изменение (*падение* [цен], *подъем* [экономики], *прирост* [производства]), обладание

[3] Литературные примеры извлечены из Национального корпуса русского языка; некоторые из них приводятся в сокращении.

(*издержки*, *расход*, *прибыль*), параметр (*занятость*, *окупаемость*, *показатель*, *объем*, *минимум*), — т. к. у них за валентностью *по* + Дат стоят, в конечном итоге, количественные изменения и значения количественных параметров. Поэтому оборот *по* + Дат при этих словах трудно назвать темой, — хотя это, несомненно, разновидность денотативной привязки (которая у количественной лексики просто имеет другой вид, чем у качественной).

В оставшемся списке лексем (более 600) с валентностью *по* + Дат выделяются следующие семантические группы:

- группа «речь» включает разные подгруппы — речевые акты: *требование*, *просьба*; интерпретации: *слухи*, *шумиха*; взаимодействие: *спор*, *диалог*; тексты (в том числе документы), а также их части: *публикация*, *договор*, *раздел*, *пункт*;
- ментальные объекты, процессы и состояния: *идея*, *мысль*, *понятие*, *предположение*, *прогноз*, *размышление, память*, *секреты*;
- интерпретация: *заслуга*, *неприятность*, *перегиб*, *перспектива*, *препятствие*, *промах*, *равенство* [не количественное], *различие*, *раскол*, *совпадение*, *уловка*, *уступка*;
- мероприятия: *заседание*, *конференция*, *конкурс*, *референдум*, *семинар*, *турнир*;
- структуры: *комиссия*, *организация*, *служба*;

Примечание. Структуры, т. е. организации, коллективы и т. д., вошли в общий список потому, что они имеют валентность, описывающую их деятельность и аналогичную пропозициональным валентностям: *комиссия по урегулированию*.

- деятельность: *забота*, *меры*, *работа*, *труд*, *усилия*, *хлопоты*, *шаги*;
- цель: *задумка*, *затея*, *план*, *программа*;
- классификаторы (металексика): *итог*, *метод*, *обстановка*, *положение*, *последствие*, *процедура*, *результат*, *ситуация*, *стандарт*, *тенденция*, *условие*, *факт*.

Мы отдаем себе отчет, что эти семантические ярлыки во многом условны — и в смысле содержания, и в смысле границ (разные лексемы могут «перетекать» в другие группы), — однако классификация абстрактной лексики — это отдельная и очень большая тема, и мы не можем углубляться в ее обсуждение.

Эти семантические группы не являются чем-то единым, а образуют своего рода шкалу удаления от прототипа к периферии. Прототипом являются слова с «настоящей» валентностью содержания и темы, т. е. с пропозициональным объектом (ментальные и речевые). Слова из этих групп нередко имеют исконное управление *о* + Пр, но наряду с ним используют *по* + Дат: *мысли о переустройстве — мысли по переустройству*; *договор о поставках — договор по поставкам*; *доклад об Ираке — доклад по Ираку*. Далее идут интерпретационные и оценочные слова, мероприятия, у которых есть денотативное содержание, а часто и тема в обычном смысле слова (*лекция по международному положению*).

У слов со значением цели (типа *план*, *программа*) содержание и тема в пределах оборота *по* + Дат, на наш взгляд, неразличимы, т. к., например, в выражении *программа по переустройству* [Х-а] содержанием является переустройство, а редуцированным способом передачи содержания, т. е. темой ('о чем программа?'), можно считать цель — и это тоже 'переустройство'. Что касается группы «деятельность» (*меры*, *работа*, *хлопоты*, *шаги*), то здесь тоже содержание и цель (которую по аналогии с предыдущей группой можно считать и темой) выражается одинаково — семантика цели не противоречит семантике содержания: *усилия / меры / шаги по преодолению кризиса* и по содержанию состоят в преодолении кризиса (т. е. каждый шаг / мера «устраняет» какой-то элемент кризиса) и в качестве цели имеют преодоление кризиса.

Примечание. Мы специально оговариваем эту особенность группы «деятельность», т. к. в работе [Иомдин 1991], где приводится подробный список значений предлога *по*, валентность содержания (*операция по удалению катаракты*) и валентность темы (*дискуссия по Прибалтике*) различаются, и случаи типа *работы по озеленению* относятся как раз к примерам валентности содержания, см. [Иомдин 1991: 109]. Поскольку для нас цель является коррелятом темы, мы считаем возможным не различать тему и содержание в тех случаях, где содержание «одноименно» цели. Кроме того, неразличению содержания и темы может способствовать эллипсис. Например, в «канцелярском жаргоне» вполне возможно эллиптическое выражение *Сегодня операции по катаракте*, где *по* + Дат допустимо трактовать как тему.

Таким образом, во всех приведенных случаях есть какая-то пропозиция, которая редуцируется до темы.

Семантические модификации абстрактных существительных в контексте *по* + Дат

Разные способы выражения темы различаются, конечно, стилистически: оборот *по* + Дат более канцелярский, книжный, чем *о* + Пр. Однако есть и семантические различия, связанные со значением предлога *по*. Исходно *по* выражает направление, маршрут, траекторию (*идти по дороге*). Не менее важно значение метафорического направления — направления деятельности (*заместитель по кадрам*).

Бóльшая «расплывчатость», неопределенность *по* по сравнению с *о* при обозначении темы связана с тем, что *по* обозначает не отдельную тему, а как бы целое направление, в рамках которого происходит деятельность: *по* + Дат — это своего рода эллипсис, сокращение, за которым могут скрываться различные аспекты содержания. Т.е. *по* обозначает самую общую тему, внутри которой возможны конкретизации: *резолюция по Ираку* может включать обсуждение вопросов о прекращении боевых действий, о выводе контингента, о гуманитарной помощи, о проведении выборов.

С другой стороны, интересно, что происходит с самими абстрактными существительными в контексте *по* + Дат.

Значение абстрактных слов вообще неопределенно и в силу этого подвижно. В результате изменения значения абстрактные слова могут мигрировать из одной группы в другую, подстраиваясь под семантику принимающей группы, приобретая соответствующую валентность (в нашем случае — валентность темы) и выражая эту валентность тем способом, который характерен для данной группы (в нашем случае — *по* + Дат). Для абстрактных слов довольно типична такая ситуация, когда в словаре некоторое значение отсутствует, но в текстах оно встречается, т. е. «несловарное» значение буквально формируется в тексте и текстом, например: *Дискуссия была направлена на обсуждение следующих вопросов: — существует ли необходимость в основных международных* ***принципах по статистике*** *достойного труда; — возможно ли создание международной базы данных с сопоставимыми показателями по достойному труду; — следует ли продолжать усилия по изучению дальнейшего использования обследований рабочей силы для сбора данных, необходимых для получения индикаторов достойного труда. В ходе ведения заседания Рабочей группы Т. Л. Горбачева высказала позицию российской делегации по представленной МОТ концепции измерения индикаторов достойного труда* [«Вопросы статистики», 2004.03.25]. Мы специально привели этот значительный по объему пример, чтобы показать, насколько распространена при абстрактной лексике валентность *по* + Дат: *принципы по*; *показатели по*; *усилия по*; *позиция по*. Но нас интересует слово *принцип*. *Принцип* в обычных значениях ‘закон’, ‘правило’, ‘воззрение’, ‘убеждение’, ‘теоретическое положение’ имеет обычно генитивное или адъективное выражение валентности (*принцип свободы совести*; *принципы международного права*; *принципы минималистской эстетики*, *демократические принципы*). В данном тексте (*международные принципы по статистике достойного труда*) *принцип* (точнее, *принципы*) употребляется в значении ‘соглашение’, ‘договоренность’, т. е. смещается в группу речи — и выражает валентность темы-содержания формой *по* + Дат. Ср. также*: 17-я МКСТ* [конференция] *приняла Руководящие* ***принципы по статистическому определению занятости*** *в неформальной экономике* [«Вопросы статистики», 2004.07.29] — конференция *приняла принципы* означает приняла документ, текст, которым человек будет руководствоваться в своей деятельности, и для такого употребления *по* + Дат — естественная форма выражения валентности темы-содержания.

Поскольку валентность содержания и темы является прототипической прежде всего для речевой и ментальной, т. е. информационной, лексики, многие абстрактные лексемы, сдвигаясь в сторону речи, мысли и — шире — информации, актуализируют информационный аспект некоторой ситуации. Например, *конференция*, *переговоры*, *семинар* и т. п. — это мероприятия. Но суть этих мероприятий — речевая и информационная: обмен информацией, получение информации путем речевого взаимодействия. Поэтому естественно, что у них есть тема. Тот же информационный аспект выходит на первый план у таких интерпретационных слов, как *раскол*, *разлад* (ср. *противоречия*) — у них тоже конструкции с *по* выражают расхождение во взглядах, т. е. относятся

к информационной сфере, ср.: *Никаких* ***противоречий по разделению*** *функций Генсовета и исполкома партии при обсуждении не возникало* [«Известия», 2003.01.14]; *Какого-то особого* ***разлада по этому вопросу*** *не было* [«Известия», 2002.07.08]; *Тем самым при внешнем согласии, в сущности,* ***раскол по проблеме*** *Ирака сохраняется* [«Военная мысль», 2004.11.15]. Ср. также другие примеры «перетекания» значений: *Есть ли у вас какие-то* ***секреты по уходу*** *за собой?* [«100% здоровья», 2003.01.15] ≈ 'приемы, методы, рецепты, рекомендации'; *А заранее избавившись от сорняков, мы во многом облегчим себе* ***труд по уходу*** *за садом* [«Сад своими руками», 2003.09.15] ≈ 'заботы, хлопоты'.

Другой важной особенностью поведения абстрактной лексики в контексте *по* + Дат является то, что многие слова с валентностью темы-содержания употребляются преимущественно или даже исключительно во множественном числе: *Их основной целью является поиск* ***путей по*** *минимизации риска возникновения ядерного конфликта* [«Зарубежное военное обозрение», 2004.05.24]; *В этом вопросе необходимо учитывать характер* ***последствий по*** *недействительной сделке* [«Арбитражный и гражданский процессы», 2004.06.28]; *Руководитель агентства железнодорожного транспорта РФ считает* ***опасения по*** *падению грузопотока неоправданными* [www.rbcdaily.ru/2008/05/14]; *Чтобы показать, что под его руководством превзошли все* ***ожидания по*** *выработке цемента, он завел много его марок* [Никита Хрущев. Воспоминания (1971)] — *последствие по* в НКРЯ не встретилось, *опасение по* — меньше 10 раз, *ожидание по* — 1 раз (на начало 2015).

В русском языке широко распространено такое явление, как лексикализация множественного числа (что часто отмечается в словарях, ср.: *вода — воды*; *выбор — выборы* и под.). Однако семантический сдвиг у формы множественного числа (далее — мн. ч.) — не просто широко распространенное явление, это активный процесс, который не всегда поддается словарной фиксации. В этой области существует множество разных вариантов. Например, *перебои*, *перегибы* употребляются преимущественно во мн. ч., хотя теоретически имеют и единственное; *перспективы*, *последствия*, употребляясь преимущественно во мн. ч., могут соответствовать единственному, ср.: *В этой фирме у него хорошие перспективы* обычно означает — есть перспектива карьерного роста. Но поскольку карьерный рост может быть связан с множеством других вещей (повышение зарплаты, увеличение числа подчиненных, личный кабинет, участие в важных решениях и т. д.), то все это как бы включается во мн. ч. *перспективы*. Т.е. здесь употребление формы мн. ч. — своего рода риторический прием, призванный показать сложность и многоплановость некоторого явления.

Наш материал показывает, что даже исходно «информационные» слова могут развивать особое значение в форме мн. ч. (с соответствующей валентностью *по* + Дат), ср.: *требование* (*требование прекратить огонь*) — это речевой акт, обращенный к конкретному лицу (лицам) и предполагающий конкретное действие, *требования* (*требования по технике безопасности*) — это какие-то правила и инструкции (нередко — письменные), обращенные к гипотетическим исполнителям в гипотетических ситуациях. Если же исходно

не информационное значение претерпевает семантический сдвиг, связанный с развитием или усилением «информационного компонента» (ср. выше *раскол*, *разлад*), то этот сдвиг часто происходит именно в форме мн. ч. Рассмотрим несколько контекстов из НКРЯ: *оправдались ожидания по доходности доллара*; *ожидания по явке избирателей полностью оправдались*; *ожидания по промышленному производству, видимо, будут скорректированы*. Если *ожидание 1* (по МАС) — это состояние по глаголу *ожидать*, то *ожидание 2* [*обычно мн. ч.* (*ожидания*, *-ий*)] — это ‘предположение, надежда’, а в нашем случае скорее — ‘прогноз’, и эта лексема с ментально-речевым значением управляет формой *по* + Дат, а управление родительным (нормальное для исходного значения, ср. *ожидание встречи*) либо приводит к семантическому искажению, ср. *ожидания доходности / явки* — бытийное значение (‘ожидается, что доходность / явка будет’), тогда как смысл выражений типа *ожидания по доходности / явке* сводится к косвенному вопросу (‘предположения о том, какая будет доходность / явка’), либо вообще невозможно, ср. ?*ожидание промышленного производства*.

Итак, анализ материала показал, что одним из вариантов денотативной привязки для нескольких классов абстрактной лексики является валентность темы (как редукция содержания). Поскольку валентность содержания / темы свойственна информационной (в широком смысле) лексике (ментальной, речевой), у других классов абстрактной лексики, приобретающих такую валентность, происходит сдвиг в сторону информационной семантики.

Что касается способа выражения валентности темы, то для нее используется оборот *по* + Дат, который является более семантически мотивированным, чем исходный для многих существительных, но семантически опустошенный генитив. Таким образом, процесс замены беспредложного управления предложным, который обычно считается одним из проявлений тенденции к аналитизму, в данном случае приводит к семантизации формы, выражающей валентность.

Литература

1. *Апресян Ю. Д. и др.* (2010) Теоретические проблемы русского синтаксиса: Взаимодействие грамматики и словаря, ЯСК, М.
2. *Иомдин Л. Л.* (1991) Словарная статья предлога ПО // Семиотика и информатика. Вып. 32. ВИНИТИ, М. С. 94–120.
3. *Курилович Е.* (1962) Проблема классификации падежей // Курилович Е. Очерки по лингвистике. М. С. 175–203.
4. *Кустова Г. И.* (в печати) Синтаксические и семантические факторы в падежных стратегиях // Труды Института русского языка им. В. В. Виноградова РАН, М.
5. *Wierzbicka A.* (1980) The case for surfase case. Karoma Publishers, Ann Arbor. Рус. пер.: Вежбицка А. Дело о поверхностном падеже // Новое в зарубежной лингвистике. Вып. 15. Прогресс, М. 1985. С. 303–340.

References

1. *Apresjan Yu. D. et al.* (2010) Theoretical problems of Russian syntax: The interaction of grammar and vocabulary, LSC, Moscow.
2. *Iomdin L. L.* (1991) Lexical entry of the word PO [Slovarnaya statya predloga PO], Semiotics and Informatics [Semiotika I Informatika], Vol. 32, pp. 94–120.
3. *Kurilovich E.* (1962) The problem of case classification [Problema klassifikatsii padezhey], in Kurilovich E. Essays in Linguistics [Ocherki po lingvistike], Moscow, pp. 175–203.
4. *Kustova G. I.* (in print) Syntactic and semantic factors in case strategies [Sintaxicheskie I semanticheskie factory v padezhnykh strategiyakh], Proceedings of the V. V. Vinogradov Russian Language Institute [Trudy Instituta russkogo yazyka im. V. V. Vinogradova RAN], Moscow.
5. *Wierzbicka A.* (1980) The case for surfase case. Karoma Publishers Inc., Ann Arbor.

АВТОМАТИЧЕСКОЕ СНЯТИЕ МОРФОЛОГИЧЕСКОЙ ОМОНИМИИ В КОРПУСАХ НОВОГРЕЧЕСКОГО ЯЗЫКА И ЯЗЫКА ИДИШ

Кузьменко Е. А. (eakuzmenko_2@edu.hse.ru),
Мустакимова Э. Г. (egmustakimova_2@edu.hse.ru)

Национальный исследовательский университет
«Высшая школа экономики», Москва, Россия

Ключевые слова: морфологический анализ, снятие омонимии, корпусная лингвистика, греческий язык, язык идиш

AUTOMATIC DISAMBIGUATION IN THE CORPORA OF MODERN GREEK AND YIDDISH

Kuzmenko E. A. (eakuzmenko_2@edu.hse.ru),
Mustakimova E. G. (egmustakimova_2@edu.hse.ru)

National Research University Higher School of Economics,
Moscow, Russia

The problem of morphological ambiguity is widely addressed in the modern NLP. Mostly ambiguity is resolved with the use of large manually-annotated corpora and machine learning. However, such methods are not always available, as good training data is not accessible for all languages. In this paper we present a method of disambiguation without gold standard corpora using several statistical models, namely, Brill algorithm (Brill 1995) and unambiguous n-grams from the automatically annotated corpus. All the methods were tested on the Corpus of Modern Greek and on the Corpus of Modern Yiddish.

As a result, more than a half of words with ambiguous analyses were disambiguated in both corpora, demonstrating high precision (>80%). Our method of morphological disambiguation demonstrates that it is possible to eliminate some of the ambiguous analyses in the corpus without specific linguistic resources, only with the use of raw data, where all possible morphological analyses for every word are indicated.

Keywords: morphological tagging, morphological disambiguation, corpus linguistics, Modern Greek, Yiddish

1. Introduction

As the usage of corpus methods becomes widespread in linguistics, the problem of ambiguity in existing corpora turns out to be more and more significant. To perform deep linguistic analysis a researcher needs language data of high quality. Meanwhile, morphological processing of the corpus data involves two steps: assigning morphological analyses to tokens and, as wordforms in a language are often ambiguous, disambiguation. Ambiguity in corpora does not allow linguists to make detailed queries and get exact results because they receive a lot of irrelevant data. Disambiguation by hand is time-consuming; therefore, it is essential that ways of computer-aided disambiguation are developed. Most of the disambiguation techniques are based on the implementation of machine learning. Machine learning implies having a huge manually disambiguated corpus, which can be used for training the algorithm. However, such resources are not available for every corpus and every language. For this reason we need to find an approach to disambiguate texts with no knowledge about the statistics of word occurrences in particular contexts and with no manual annotation. Despite these rough demands, they should show high accuracy and amplitude.

In this paper we consider automatic disambiguation techniques for the corpora of Modern Greek and Yiddish, which do not have pre-disambiguated subcorpora. We combine several existing disambiguation algorithms in a more effective way, adapt POS-tagging algorithms to the disambiguation problems (Brill 1995) and develop a technique of our own (disambiguation on the basis of unambiguous n-grams found in the corpus). We estimate the effectiveness of each technique and compare them.

The originality of our work lies in absence of any manually processed data. Moreover, we work with morphologically rich languages. All the data available to us is raw corpora in which every word is assigned all possible morphological analyses.

2. Corpora

2.1. The corpus of Modern Greek

The corpus of Modern Greek[1] consists of 26 million tokens. The majority of texts come from Greek newspapers and belong to the 21st century. Also there are such genres as fiction (both native and translated works), poetry, publicistic writing and scientific literature. These texts belong to 19th–21st centuries. The Corpus of Modern Greek is based on the EANC platform (Arkhangelskiy et al. 2013). Morphological information in this corpus is stored according to the UniParser standard.

Every word is assigned all possible analyses; for example, the word occurence μέσα could be assigned the following analyses:

1. μέσα, ADV, “inside”;
2. μέσο, NOUN,n,pl,acc, “medium”;
3. μέσο, NOUN,n,pl,nom, “medium”;

[1] http://web-corpora.net/GreekCorpus/search/

4. μέσος, ADJ,pos,n,pl,acc, “middle”;
5. μέσος, ADJ,pos,n,pl,nom, “middle”.

Before performing disambiguation, we estimated baseline parameters of ambiguity in our corpus (Table 1):

Table 1. Baseline parameters for ambiguity in the corpus

Number of tokens	Percentage of ambiguous words	Ambiguity rate
26,075,298	43%	1.64

In this table the parameters signify the following:

- Number of tokens—number of words in the corpus
- Percentage of ambiguous words—the ratio of tokens which have more than one analysis to the overall number of tokens in the corpus
- Ambiguity rate—the ratio of all tags in the corpus to all tokens

Most of the words had 2 or 3 analyses, and sometimes they had 4 or even 5 analyses.

Overall, there were almost 10 thousand (9,987, to be exact) different types of ambiguity, and there were 11 thousand (10,842) different ambiguous word instances.

The most frequent types of ambiguity were the following (different morphological analyses are separated with dashes):

1. το,ART,n,sg,acc—το,ART,n,sg,nom—το,PRO,n,sg,acc—τον,ART,m,sg,acc
2. του,ART,m,sg,gen—του,ART,n,sg,gen—του,PRO,m,sg,gen—του,PRO,n,sg,gen
3. του,ART,m,sg,gen—του,ART,n,sg,gen
4. είμαι,V,pres,3,pl—είμαι,V,pres,3,sg
5. με,PR—με,PRO,1p,sg,acc
6. της,ART,f,sg,gen—της,PRO,f,sg,gen
7. των,ART,pl,gen—των,PRO,pl,gen
8. την,ART,f,sg,acc—την,PRO,f,sg,acc
9. είμαι,V,past,3,pl—είμαι,V,past,3,sg
10. τα,ART,n,pl,acc—τα,ART,n,pl,nom—τα,PRO,n,pl,acc

These 10 types of ambiguity out of 10 thousand overall together constitute 15% of ambiguity in the corpus.

2.2. The corpus of Modern Yiddish

The Corpus of Modern Yiddish[2] (CMY) is a joint project of the Russian Academy of Sciences and the University of Regensburg, which started in 2007. The corpus comprises mainly publicistic texts, fiction is represented to a much lesser degree. For now the volume of the CMY is about 4 million tokens.

[2] http://web-corpora.net/YNC/search/

As in the Corpus of Modern Greek, each word in the CMY is supplied with a list of all possible morphological interpretations. The ambiguity rate in Yiddish is even higher than in Greek. Baseline parameters for CMY are shown in Table 2.

Table 2. Baseline parameters for ambiguity in the Corpus of Modern Yiddish

Number of tokens	Percentage of ambiguous words	Ambiguity rate
4,144,524	39,5%	2.026

In the case of CMY it is impossible to resolve all cases of ambiguity. The first reason is that almost all nouns are supplied with at least 4 analyses. A noun in Yiddish has 4 cases (nominative, genitive, dative, accusative), but the case forms look identical for most nouns. Since complete resolution for all nouns is impossible and partial resolution would result in inconsistent markup and inconvenient corpus search, such cases of ambiguity will be ignored. The second reason is that verbs can merge with pronouns into one word and dative case prepositions merge with definite articles. Such merges are supplied with at least two tags. Thus, despite the fact that we want to map each token in the corpus to a single morphological interpretation, we have to accept multiple analyses for nouns and merged wordforms.

The corpus has 729 types of different combinations of tags in ambiguous words and about 24,000 different words that are homonymous. Observe that the corpus has about 2 million ambiguous words in total, and only 24 thousand different ambiguous words. According to Zipf's Law and these numbers, it is logical to assume that resolving some small amount of the most frequent homonymy types should significantly lessen the amount of ambiguity.

3. Related work

We are not the first to apply data-driven algorithms to the task of morphological disambiguation. It has been already done for such languages as Icelandic, Swedish and Turkish. The researchers working on these languages employed the Brill algorithm, and so did we. Similarly to our decision, this algorithm is not applied solely, but in combination with other approaches, such as composing linguistic rules and using n-grams. The results for other languages are the following: for Icelandic the precision of 93.65% was achieved (Helgadóttir 2004). For Swedish the results are slightly worse—only 84.5% precision (Maurier et al. 2003). For Turkish, on the contrary, significant results are reported—the authors managed to achieve the precision of 96.8% (Sak et al. 2007). Maybe this result is due to the morphology of Turkic languages, which is more easily formalized compared to languages with cumulative morphology.

Some research has been done specifically for the Greek language. However, corpora of the Greek language are not numerous: there are such corpora as HNC (Hatzigeorgiu et al. 2000), DELOS (Kermanidis et al 2002), and CGT (Goutsos 2010). Meanwhile, these corpora do not provide morphological disambiguation, or it is of poor quality. There were also some attempts to design tools for disambiguation in the Greek

language, for example, the research described in (Petasis 1999). However, in this case the tagset is very limited, so no detailed morphological information is provided. Also this approach uses a pre-disambiguated part of the corpus that serves as a golden standard. Therefore, our work is very different from the previous research because we, as it was already stated, do not use manually processed data.

If we talk about Yiddish, there are three written corpora of Yiddish: the Aston corpus of Soviet Yiddish which does not have morphological annotation, Yiddish Treebank of the University of Pennsylvania (no one knows how exactly it was annotated, probably manually) and the CMY, which is a comprehensive, annotated and a freely available corpus. Also the Yiddish language lacks tools for disambiguation, so we can not compare the result of the task with the works of previous researchers.

As we can see, Greek and Yiddish can be called under-resourced languages to the full extent: there are not many corpora for these languages and disambiguation in these corpora was not properly performed. This means that Greek and Yiddish need a method for disambiguation which would not require linguistic resources and will provide high quality despite these constraints.

4. Methods

Our decision was to find the way to combine data-driven and rule-based algorithms[3]. We used the following data-driven methods:

- trasformation-based error-driven learning (Brill 1995a; Brill 1995b);
- using data about bigrams and trigrams in which the word under consideration can be found;
- the user interface for disambiguating based on bigrams and trigrams.
  Also we used the hand-crafted rules approach.

For evaluation of the methods we used a testing part of the Greek Corpus which contains 866,091 tokens. In the case of Yiddish we used the whole CMY as a test corpus since its volume is fairly small.

4.1. The Brill algorithm

Transformation-based error-driven algorithm for pos-tagging and disambiguation purposes was developed by Eric Brill in 1995. This algorithm is very useful in our situation because it is unsupervised (which means that we do not need the disambiguated corpus). We can achieve significant results by just using unambiguous word instances from our corpus.

We tested two versions of Brill disambiguation algorithm:

---

[3] In (Halperen et al. 2001) it was shown that taggers combining several approaches result in a higher accuracy.

1. The version that executed only disambiguation with respect to the part of speech (the cases where words had analyses with different POS-tags were resolved)
2. The version that executed full disambiguation (the cases where words had analyses with the same POS-tag, but different values, were also resolved)

After applying the first version of the Brill algorithm to the test corpus the ambiguity parameters changed in the following way (Table 3):

Table 3. Ambiguity parameters of the test corpus after POS-disambiguatuion by the Brill algorithm

Corpus	Number of tokens	Percentage of ambiguous words	Ambiguity rate
Greek	866,091	29.0%	1.36
Yiddish	4,144,524	35.1%	1.86

Also we applied the second version of the Brill algorithm to the testing corpus of Greek, and the ambiguity parameters changed in the following way (Table 4):

Table 4. Ambiguity parameters of the testing corpus after full disambiguation by the Brill algorithm

Corpus	Number of tokens	Percentage of ambiguous words	Ambiguity rate
Greek	866,091	37%	1.57

In the case of Yiddish, the quality of POS-disambiguation performed by the Brill algorithm was fair enough. However, the picture for the Greek language was different: the first version of the Brill algorithm works extensively (recall ~41.4%), but most words are changed incorrectly (precision ~8.2%). The second version of the algorithm, however, shows high precision (~74%), but changes very few words (recall ~8.7%). This results in similar values of F_1-score (13.73 in the first case and 15.62 in the second case). The similarity of the scores for the methods shows that they are almost equally effective (or, in our case, ineffective).

Then we tested the results of the algorithm when we first applied the Brill algorithm in the full mode, and then finished disambiguation by the POS-version. The idea was that after applying the first variant of the algorithm the number of unambiguous words increases and drives the second version to be more accurate. The results are displayed in Table 5:

Table 5. Ambiguity parameters and effectiveness measures after applying two versions of the Brill algorithm

Corpus	Number of tokens	Percentage of ambiguous words	Ambiguity rate	Precision (%)	Recall (%)	F_1-score
Greek	866,091	26%	1.32	22.7	49	31.02

As we can see from the table above, POS-disambiguation by the Brill algorithm indeed becomes more effective when applied to the pre-processed data. Its precision increases, though it is still on the low level, and the F_1-score shows that such results are more valuable than the previous results (31.02 compared to previous ~14). This experiment shows that POS-disambiguation by the Brill algorithm can serve as the final step in the process of disambiguation, when it would become more effective.

4.2. Bigrams and trigrams

This approach is similar to probabilistic Markov models described in (Kupiec 1992), but it works in a different and simpler way. Training a good bigram model requires a manually annotated corpus which we do not have. For this reason, we decide to automatically extract non-ambiguous parts of the corpus and treat it as an etalon. Non-ambiguous bigrams are those both words of which have no ambiguity.

In the Python programming language, the bigram model is realized as a dictionary where each key is a morphological tag and the corresponding value is an array of tuples. Each tuple contains a) a tag that may follow the key and b) the probability of a bigram (key + such tag). Then we use a script that runs through the corpus and looks for ambiguous words. For each such word the script checks whether the previous word is not ambiguous, and if so the script would consult the bigram model, choose the most probable tag out of the given and delete all the redundant analyses from the current word interpretation.

Let us illustrate how the model works. For example, the model contains a frequent nonambiguous bigram *N,m,pl* followed by *V,pres,pl,1*. When the script meets the tag *N,m,pl* followed by an ambiguous word with tags *V,pres,pl,1_V,pres,pl,3_V,inf,* it would keep *V,pres,pl,1* and delete the others.

This algorithm does not use any other statistics about contexts in which particular word analyses can be met, so its results can be erroneous. Surprisingly, the accuracy of this method was rather sufficient, and we will demonstrate it further.

We applied the algorithm to the testing corpus and received the following results for the ambiguity parameters and the effectiveness of the algorithm (Table 6):

Table 6. Ambiguity parameters of the testing corpus after applying the bigrams algorithm and the effectiveness of the algorithm

Corpus	Number of tokens	Percentage of ambiguous words	Ambiguity rate	Precision (%)	Recall (%)	F_1-score
Greek	866,091	38%	1.59	83	8.1	14.82
Yiddish	4,144,524	24.4%	1.65	78	—	—

As we can see from the table, this simplified and easy to execute model, surprisingly, demonstrates the same level of effectiveness as the intelligent data-driven Brill algorithm and even has the higher level of precision.

4.3. An interface for disambiguation by hand

All the approaches considered above are more or less effective, but all of them make mistakes. Every method considered earlier generated incorrect changes of tags, so that the correct tag for a particular word was deleted. Manual disambiguation is usually more accurate, but it is a very tedious process.

Imagine that a corpus has 3,000 instances of a bigram *ART, m, sg + N, m, sg_N, m, pl*, where the second word is ambiguous and has two possible tags. The correct tag is obvious, but a human would have to disambiguate this one simple bigram 3,000 times. It would be more convenient to resolve such morphological ambiguity just once and then automatically apply the result to all corresponding cases. For such disambiguation process we designed a program which interacts with a linguist and works as an automatic text processor.

The program collects ambiguous unigrams, bigrams and trigrams in the corpus and sorts them by frequency. Then the program shows one of the collected items to the user and offers to choose which variant is correct. The user can mark the correct answer or delete the wrong variants. If the user is not sure how to resolve ambiguities, they can be skipped. The accuracy of this method depends on the knowledge of language. Assuming that the linguist knows the language, this algorithm is very accurate.

This method stands closer to the rule-based methods as it does not depend on the data—the user can choose the right variant even when all the words in a bigram or trigram are ambiguous. Therefore, this method can be the first step in the process of disambiguation because its results cannot significantly change with the increase of unambiguous words in the corpus. In contrast, this method can supply data-driven methods with the higher number of unambiguous contexts and consequently improve their precision and recall while itself demonstrating supposingly high precision.

We received the following results for this user-guided disambiguation (Table 7):

Table 7. Ambiguity parameters of the testing corpus after applying user-guided disambiguation and the effectiveness measures for the algorithm

Corpus	Number of tokens	Percentage of ambiguous words	Ambiguity rate	Precision (%)	Recall (%)	F_1-score
Greek	866,091	31.8	1.50	84.8	31.1	45.51
Yiddish	4,144,524	34.3	1.88	97.5	—	—

As we can see, this method turned out to be very effective as it demonstrated both high recall and high precision. Actually, the value of precision is not equal to 100% because in some cases incorrect tags were deleted, but the word still had more than one analysis, which was counted as an imprecise case. In fact, this method did not generate incorrect tag changes, in contrast to the previous data-driven methods.

5. Results and discussion

We gave an overview of our methods and their effectiveness when they were applied to the training corpus. However, as it was mentioned earlier, the disambiguation becomes more accurate and extensive when several methods are combined.

The key point of the combination of methods is that data-driven methods work better if they are given more positive material. The more unambiguous data we have the more precise the method is and the higher the recall percentage is. Therefore, our aim is to use the methods which are not data-driven first, then to use methods which are more precise so that they could create more positive data for methods which are less precise.

All in all, we chose the following order:

1. user-guided disambiguation, which has high precision and is not data-driven;
2. the bigrams algorithm, which is data-driven, but with high precision;
3. the Brill algorithm (full disambiguation, not only by POS-tags), which has lesser precision;
4. rule-based disambiguator for ambiguity types left (for Greek).

Table 8 demonstrates the changes in process when we applied to the training corpus our disambiguation methods in this order:

Table 8. Ambiguity parameters and effectiveness measures for the combinations of methods

Corpus	Method	Percentage of ambiguous words	Ambiguity rate	Precision (%)	Recall (%)
Greek	user-guided	31.8%	1.50	84.80	31.10
	+ bigrams	29.0%	1.45	85.00	36.00
	+ Brill	26.0%	1.40	79.92	43.48
	+ rules	23.0%	1.35	82.41	50.60
Yiddish	user-guided	34.3%	1.89	97.50	—
	+bigrams	17.8%	1.48	82.70	—
	+ Brill	15.2%	1.39	81.70	—

As we can see from this table, the ambiguity rate gradually falls down with every method, and recall rises while precision stays on the high level. All this shows that every method indeed becomes effective if applied in the right combination with other methods.

In this paper, we have considered different disambiguation methods for the case when machine learning and supervised methods based on the pre-disambiguated corpus are not accessible to the researcher. We adapted several data-driven approaches such as the Brill algorithm and the Viterbi algorithm so that they became useful in our situation. Also we designed several techniques of our own such as user-guided disambiguation by bigrams and trigrams and supported our scheme with a conventional rule-based parser. In the end, we managed to resolve a significant number of ambiguous analyses in our corpora and proved that it could be done without using specific linguistic resources, such as training corpora disambiguated by hand.

Of course, it can be argued that the precision and recall we managed to achieve are not high enough to suit the needs of linguistic research. However, in the situation of the total absence of disambiguation tools for these languages developing approaches to disambiguation is vital, and, as the research concerning Swedish, Turkish and Icelandic shows, the quality of disambiguation can be improved, so we plan to adapt the solutions proposed for other languages with respect to disambiguation.

References

1. *Arkhangelskiy, T., Belyaev, O., & Vydrin, A.* (2012). The creation of large-scaled annotated corpora of minority languages using UniParser and the EANC platform. Proceedings of COLING 2012: Posters. Mumbai: The COLING 2012 Organizing Committee, 2012. Ch. 9. P. 83–91.
2. *Brill E.* (1995a). Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. Computational linguistics, 1995, V. 21, № 4, pp. 543–565.
3. *Brill E.* (1995b). Unsupervised learning of disambiguation rules for part of speech tagging. Proceedings of the third workshop on very large corpora. Somerset, New Jersey: Association for Computational Linguistics, 1995, V. 30, pp. 1–13.
4. *Goutsos, D.* (2010). The corpus of Greek texts: a reference corpus for Modern Greek. Corpora, 5 (1), 29–44.
5. *Van Halteren H., Daelemans W., Zavrel J.* (2001). Improving accuracy in word class tagging through the combination of machine learning systems. Computational linguistics, 2001, V. 27, № 2, pp. 199–229.
6. *Hatzigeorgiu, N., Gavrilidou, M., Piperidis, S., Carayannis, G., Papakostopoulou, A., Helgadóttir, S.* (2004). Testing data-driven learning algorithms for pos tagging of icelandic. Nordisk sprogteknologi, 2000–2004.
7. *Kermanidis, K. L., Fakotakis, N., & Kokkinakis, G.* (2002). DELOS: An Automatically Tagged Economic Corpus for Modern Greek. In Proceeding of LREC 2002. P. 718–722.
8. *Kupiec J.* (1992). Robust part-of-speech tagging using a hidden Markov model. Computer Speech & Language, 1992, V. 6, № 3, pp. 225–242.
9. *Marier, F., & Sjödin, B.* (2003). A part-of-speech tagger for Swedish using the Brill transformation-based learning. Projektarbeten 2003, 102.
10. *Sak, H., Güngör, T., & Saraçlar, M.* (2007). Morphological disambiguation of Turkish text with perceptron algorithm. In Computational Linguistics and Intelligent Text Processing (pp. 107–118). Springer Berlin Heidelberg.
11. *Spiliotopoulou, A., & Demiros, I.* (2000). Design and Implementation of the Online ILSP Greek Corpus. In Proceeding of LREC 2000.

AUTOMATIC CLASSIFICATION OF WEB TEXTS USING FUNCTIONAL TEXT DIMENSIONS

Lagutin M. B. (lagutinmb@mail.ru)*,[1],
Katinskaya A. Y. (a.katinsky@gmail.com)[2],
Selegey V. P. (Vladimir_S@abbyy.com)[2,3,4],
Sharoff S. (s.sharoff@leeds.ac.uk)[2,5],
Sorokin A. A. (alexey.sorokin@list.ru)*,[1,2,3]

[1]Lomonosov Moscow State University, Moscow, Russia
[2]Russian State University of Humanities, Moscow, Russia
[3]Moscow Institute of Physics and Technology, Moscow, Russia
[4]ABBYY, Moscow, Russia
[5]Leeds University, Leeds, UK

The work addresses automatic genre classification of Web texts. We show that functional text dimensions could be used for this tasks, with their stable combinations (clusters) corresponding to genres. Basing on a gold standard corpus, we construct a list of such genres. We also show that functional dimensions values can be automatically extracted from language features. In the conclusion we discuss the application of our results for automatic annotation of large Web corpora.

Introduction

It is well-known that an additional genre annotation could be very useful for corpus studies. Genre obviously affects lexical, syntactical and other text parameters. Using of genre information seems promising in different tasks of computational linguistic, e.g. for language models refinement. Therefore a reliable genre annotation is very desirable and ideally it should be done automatically since manual annotation even of medium-size corpora is an extremely labour-intensive task. It is especially true for corpora of texts from the Internet, in particular for General Internet-Corpora of Russian Language (GICR) which is currently under development (Belikov et al., 2012).

Standard systems cannot fit genre structure of Internet perfectly since they lack such concepts as “blog” or “forum discussion”. Mechanical extension of the system with new categories does not correct this deficiency: numerous annotation experiments have shown that inter-annotator agreement for Internet texts is rather decent. It is not due to a bad qualification of annotators or unclear instructions: most Internet texts demonstrate a real mixture of various genres without any clear border between them. That leads to an objective disagreement between annotators and low quality of automatic genre classification.

* Corresponding author

One possible way to avoid this problem is to replace genres with Functional Text Dimensions (FTDs). The system of FTDs of S. Sharoff (Forsyth and Sharoff, 2014) showed better inter-rater agreement both for Russian and English languages than genres of D. Biber (Egbert and Biber, 2013). The drawback of this approach is that assigned rates are uninformative: it is not very clear, for example, what means that a particular text has a value of 1 for the feature A7. Therefore, there is a need to establish the correspondence between the genre of a text and its FTDs.. Since the dimensions are not independent, some combinations of FTDs are more stable than others. Such stable and frequent combinations form analogues of traditional genres. Hence clustering of FTD values is an unavoidable preliminary stage of genre classification for web texts.

The present work extends our previous study (Sorokin, Katinskaya, Sharoff, 2014) addressing similar questions. We discovered several natural and stable clusters in the space of FTDs, however, their detection suffers from a serious noise. The noise originates both from imperfect annotation and an unbalanced corpus structure. We have tried to improve both the homogeneity of the corpus and the reliability of its annotation. We also made a preliminary experiment on automatic FTD detection. The achieved percentage was quite high (about 70%) which gives us hope for the further automatic genre annotation of the whole corpus or at least a large segment of it. We start the paper with describing our corpus and its preprocessing. Afterwards we make a statistical analysis of the FTD space and describe the algorithm of automatic genre classification. In the conclusion we discuss how to use our method for creating automatic genre annotation.

Corpus and functional dimensions

Our word is devoted to automatic genre classification. Since, in the strict sense, the notion of genre is not defined for texts from the Web, beforehand we analyze the space of Functional Text Dimensions (Forsyth, Sharoff, 2014) and how texts are located in this space. We used 17 FTDs given in Appendix I. In our previous studies we picked out a benchmark corpus of 618 texts. It contains most of the texts from our previous study (Sorokin, Katinskaya, Sharoff, 2014) except for the ones which do not permit reliable annotation by the FTDs. The corpus was enlarged by 90 texts from each of three popular platforms: blogs.mail.ru, vkontakte.ru and livejournal.com (its Russian segment). When collecting this corpus, we tried to cover as much various combinations of FTD values as possible. Each text was carefully annotated by two raters by 17 FTDs. Annotation scheme and guidelines were thoroughly examined during our previous studies, so we tried to make the annotation process as objective as possible. The presence of each dimension was rated on the following scale:

- 0—absent;
- 0,5—slightly;
- 1—partially;
- 2—present at most part.

The inter-rater agreement (Krippendorff α) achieved 90–95%, which is unattainable when using traditional genre systems. The annotation procedure was extensively tested in our previous research, so its results form a "gold standard" for future annotation studies.

Statistical analysis of annotation results

The intermediate goal of our study is to reveal stable combinations (or clusters) of FTD values ("pseudogenres"). Let us analyze the annotation results using the histogram below. The distribution of FTD values vary between dimensions. The features **A4**, **A5**, **A7**, **A18** are more rare than others and **A19** was detected only in 13 texts. So there is little hope to find any clusters in the subspace of these features. Even such clusters been found, their size would be too small to allow reliable automatic detection.

On the opposite side, the dimension **A16** is the most frequent. Usually it is found together with other FTDs. The explanation is abundance of encyclopedic texts in our corpus. The features also vary by the fraction of '2' scores: the dimensions **A4**, **A7**, **A9**, **A12**, **A14** were scored very categorically with the fraction of '2'-s above 70%. On the contrary, the annotators were not so confident in assigning features **A1**, **A3**, **A6**, **A15**, **A17**, **A19** (30–40%). The extreme uncertainty was **A17** ("evaluation") with only 22% of '2' between positive values.

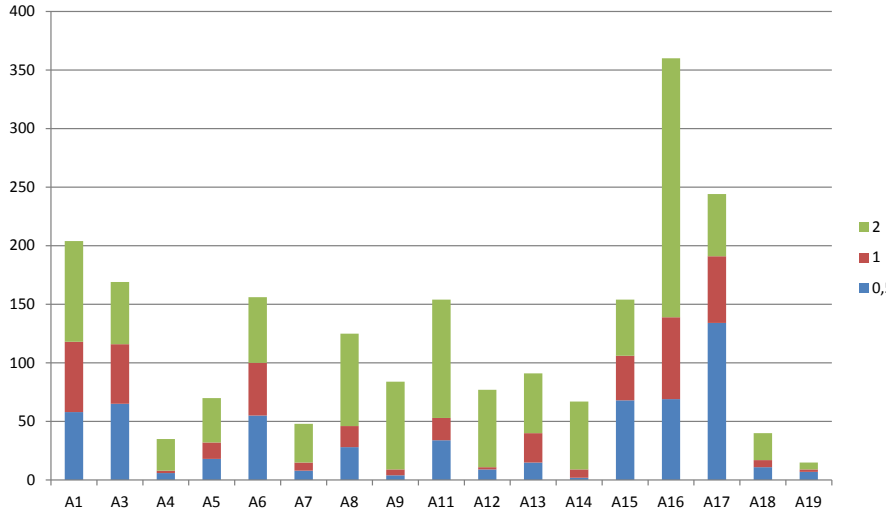


Figure 1. The distribution of FTD values

Geometrically, the annotations are the points of 17-dimensional cube located on the grid with 0.5 interval. We are looking for clusters in this cube defining a cluster as a dense and isolated subset of points. Since borderline texts are ubiquitous, there is no hope for isolation, so we care only about density. Formally, our task is to find the vertexes of 17-dimensional cube with the densest clouds of points around them. Such points will be the centers of clusters.

Clusterization: method and results

To facilitate clustering we binarize the values of the FTDs. Extreme values 0 and 2 are naturally mapped to themselves. Since the difference between '0' (absent) and '0.5'

(slight) is rather vague, we also map 0.5 to 0. Processing 1s in the same manner would be a crude oversimplification, so we use a twofold recoding. We create two datasets, the first one with 1s replaced by 0s, and the second with 1s replaced by 2s. We perform clustering for both datasets and then compare the results. By clustering in this context we mean calculating the frequencies of cube vertexes and detecting the most stable combinations.

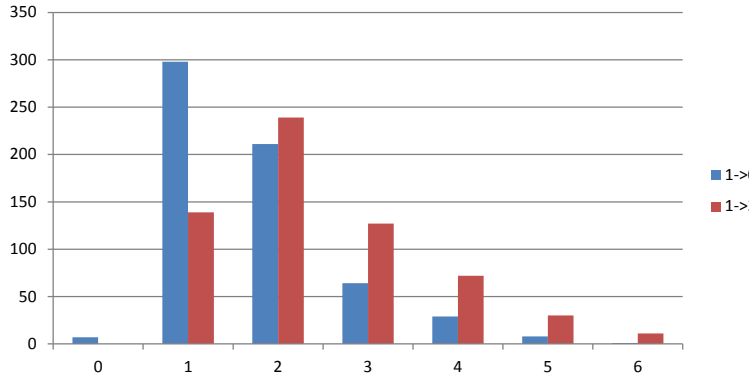


Figure 2. Histogram of the number of positive features for FTD combinations

For a vector of FTD values we call its rank the number of positive features in that vector. Figure 2 shows the frequency distribution of dataset points after various recodings. After 1→0 recoding the most frequent rank is 1 (298 times) and for 1→2 recoding such rank is 2 (239 times). Weights greater than 4 are rare: there are 5 such vertexes in the case of 1→0 recoding and 41 in case of 1→2 transformation. Both this quantities are negligible as compared to the size of the whole corpus so we may restrict our attention to combinations with 4 or less nonzero values. That gives us 3214 possible points in 17-dimensional cube.

Table 1 shows 24 top vertexes according to their frequencies after 1→0 and 1→2 recodings. For every prototype (cluster center) **v** we also measure **CW(v)**—the number of initial vertexes (before recoding) for which v is the closest prototype between the 24 selected with respect to standard Euclidean distance. In case of ties the frequency is divided by the number of closest prototypes. The prototypes are sorted in the descending order by their **CW**. Most of the cluster centers are of rank 1 and 2 except for the two prototypes of higher rank: **A14+A15+A16** and **A3+A6+A11+A17**. For the majority (352) of annotation vectors the number of closest prototypes is 1. In 92 cases the point had 2 prototypes on the same distance (most of the time the distance of 1). In 169 cases (27,3%) the points had no prototype on the distance less than 2 and were considered as noise.

Table 1. The most frequent combinations of FTD values

A1	A3	A4	A5	A6	A7	A8	A9	A11	A12	A13	A14	A15	A16	A17	A18	A19	Weight1->0	Weight1->2	CW	Rank
0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	71	53	68.20	1
0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	75	37	59.75	1
0	0	0	0	0	0	0	2	0	0	0	0	0	2	0	0	0	37	46	42.00	2
0	0	0	0	0	0	0	0	0	2	0	0	0	2	0	0	0	25	29	31.00	2
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	33	10	26.15	1
0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	30	20	26.00	1
0	0	0	0	0	0	0	0	0	0	0	2	0	2	0	0	0	24	22	25.50	2
0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	21	4	17.50	1
0	0	0	0	0	0	0	0	0	0	0	2	2	2	0	0	0	12	19	17.00	3
2	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	11	14	16.00	2
0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	13	6	13.00	1
0	0	0	0	0	0	0	0	0	0	0	0	2	2	0	0	0	7	15	11.25	2
0	2	0	0	2	0	0	0	2	0	0	0	0	0	2	0	0	6	4	11.00	4
0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	14	2	10.70	1
0	0	0	0	0	0	0	0	0	0	0	0	0	2	2	0	0	3	10	10.50	2
0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	16	2	10.00	1
0	0	0	0	0	0	0	0	0	0	0	2	2	0	0	0	0	10	5	9.00	2
2	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	8	3	9.00	2
0	0	0	2	0	0	0	0	0	0	0	0	0	2	0	0	0	5	3	8.50	2
0	0	0	0	0	0	0	0	2	0	0	0	0	0	2	0	0	7	1	8.00	2
2	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	11	6.70	2
0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	7	1	5.40	1
0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	7	3	5.00	1
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	7	0	1.15	0

Clusters and their classes

Most of the elicited clusters are too small to be used as pseudogenres: only 2 clusters from 24 are of size 60 or greater (>10% of the collection) and 5 more clusters contain 25–40 texts (5–7%).Hence reliable automatic classification for these clusters is a hopeless task since no algorithm can detect such small classes. Fortunately, the clusters themselves possess hierarchical structure and can be grouped into higher order classes. When grouping the clusters we used the correlation between the features **A1–A18** and the diagram of cluster proximity (Figure 3). The sizes of the rectangles are proportional to the size of the corresponding cluster. Note that the dimension **A19** is excluded from the future consideration due to its scarcity.

We made the following decisions on the basis of this diagram:

1) The most inhabited clusters **A8** ("news texts") and **A16** ("encyclopedic texts") form single classes.
2) Any cluster containing the FTDs **A14**, **A15** belong to the same class due to a high weight of clusters **A14+A15+A16** and **A14+A16**, and considerable correlation between **A14** and **A15** ($\rho = 0{,}42$). Note that the features **A14** ("scientific character") and **A15** ("texts to specialists")' have a lot in common already by their definition.
3) The clusters containing **A3**, **A6**, **A11** and **A17** are joined together since this features rarely appear severally and their correlation (ρ(**A3**, **A6**) = 0,5; ρ(**A3**, **A11**) = 0,54; ρ(**A3**, **A17**) = 0,4) is high (e.g., significant on $\rho = 0{,}001$ level).

4) Since the weight of cluster **A1+A13** is high (16), and their correlation ρ(**A1**, **A13**) = 0,32 is significant on the level $\rho < 0{,}0001$, we unite all the clusters containing **A1** and **A13** (except for **A1+A11** which is already in the class for **A11** dimension).
5) Join the clusters **A9**, **A9+A16** together, as well as **A12** and **A12+A16**.
6) The features **A4**, **A5**, **A7**, **A18** do not have enough positive values to form classes but their total weight (104 texts) is too high to consider them as noise. We sort these features according to their frequencies (from high to low) and attribute yet unannotated text to the first FTD with value of 1 or 2. For example, the FTD **A7** ("instructive text") occurred in this sense in 36 texts. Remaining dimensions are tightly correlated (ρ(**A4**, **A18**) = 0,44; ρ(**A4**, **A5**) = 0,28) and it is natural to join them together. Certainly, these decisions are justified only for this collection. We call this method the method of presence.

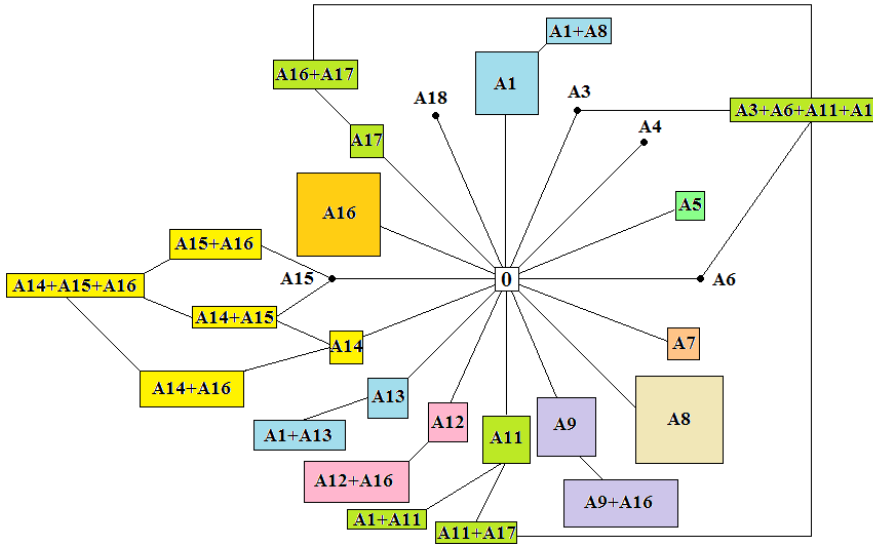


Figure 3. Diagram of cluster proximity

We obtain a list of 9 clusters **C1**–**C9** with their rounded weights as superscripts:

Table 2. Joining clusters to classes

Class	Clusters in the class
C1	A1[26], (A1+A13)[17], A13[11], (A1+A8)[7]
C2	A11[18], (A3+A6+A11+A17)[11], (A11+A1)[9], (A11+A17)[8]
C3	A8[68]
C4	(A9+A16)[42], A9[26]
C5	(A12+A16)[31], A12[10]

Class	Clusters in the class
C6	(A14+A16)[26], (A14+A15+A16)[17], (A15+A16)[11], (A14+A15)[9], A14[5]
C7	A16[60]
C8	A7[35] (*weight based on the presence method*)
C9	(A4 or A5 or A18)[69] (*weight based on the presence method)*

86 of 618 texts (13,9%) were not attributed to any class and were considered as noise. For comparison we give the results of clusterization in our previous work.

Table 3. Clusters of FTDs according to [Sorokin et al., 2014]

Class	Principal dimensions	Size
"instructive texts"	A7	21
"news texts"	A8	64
"legal texts"	A9	11
"scientific or technical texts"	A14, A15	13
"encyclopedic texts"	A16	49
"advertising texts"	A1, A12, A17	13
"propaganda texts"	A1, A13, A17	13
"noise"	—	131

Thus, our results roughly correspond to our previous work. Slight differences arose from corpus modifications and changing of the clustering procedure. It is worth noting that whereas clusterization of the present work is in some sense "manual" and "ad hoc", in (Sorokin, Katinskaya, Sharoff, 2014) we used a fully automatic procedure. The compatibility of results justifies the claim that clusters are linguistically relevant pseudogenres, not just the mathematical curiosity, and correspond to some observable language features. In what follows we study this problem for some predefined set of language features.

Language features

To address automatic genre classification we use the feature space of 40 language features **B1–B40** given in Appendix 2. The inventory of features was built for the experiment with applying multi-dimensional analysis (see Biber, 1988) to Russian and was based on the set of English linguistic features presented in (Biber et al, 2007). We adapted the features for Russian grammar and slightly restricted their set basing on the accessibility of features for automatic extraction without involving external complicated tools. We calculated the absolute frequencies of features using special program taking the morphologically tagged text as input. All the counts except for word length, sentence length and type/token ratio were normalized by the text length. As a result we obtained a 618x40 matrix containing the frequencies of 40 linguistic variables for each text. Before dealing with automatic classification we study the distribution of these features. Figure 4 contains the scatter diagram for the first 5 features.

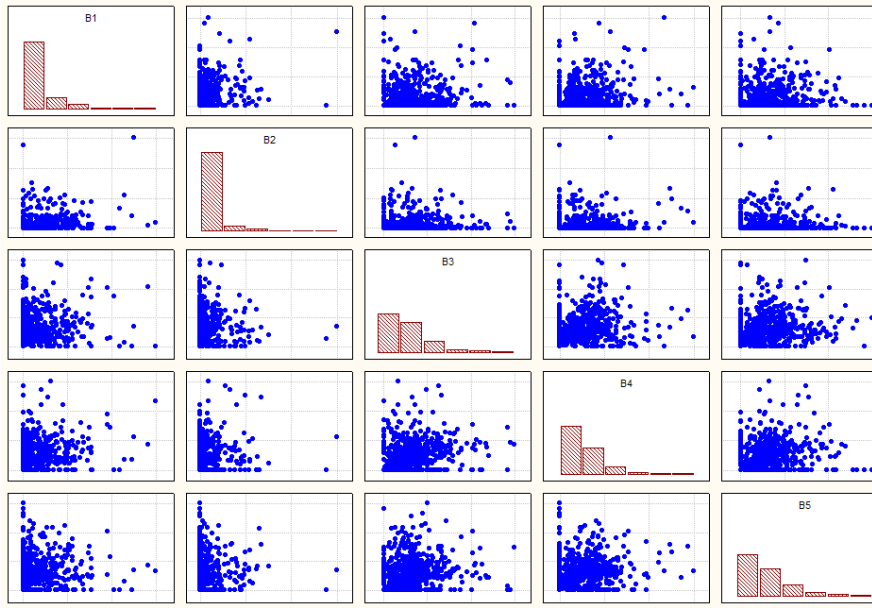


Figure 4. Scatter diagram for the features **B1–B5**

Analysis of Figure 4 leads to the following observations:

a) there are some outliers (for example, for the feature **B2**);

b) zero values are rather frequent, about 24% of all texts;

c) the distribution tails are quite heavy (see the histograms).

To prevent the distortion of regression coefficients by outliers we replace them by the closest "inlying" value. Heavy distribution tails make the model unrobust; we applied the Box-Cox normalizing transformation (Kutner et al., 2004) to reduce their effect. The value B is mapped to a new value B′

$$\mathbf{B}' = \frac{\mathbf{B}^{\lambda} - 1}{\lambda}$$

with λ estimated automatically using maximal likelihood.

We used the following scheme of transformation:

a) all zeros were temporarily referred as missing values;

b) the Box-Cox transformation was applied to the features with missing values;

c) the features were standardized to have mean 0 and deviation 1.

$$\mathbf{X} = \frac{\mathbf{B}' - \mu}{\sigma}$$

After this transformation the absolute values of the coefficients correspond to their effect

d) the missing values were replaced by -3 (an actual minimal value of a standardized normal distribution according to "3 sigma rule").

The scatter diagram for transformed features is given on Figure 5. The tails became less heavy and the shape better fits two-dimensional normal distribution, which makes the estimation of regression coefficients more robust.

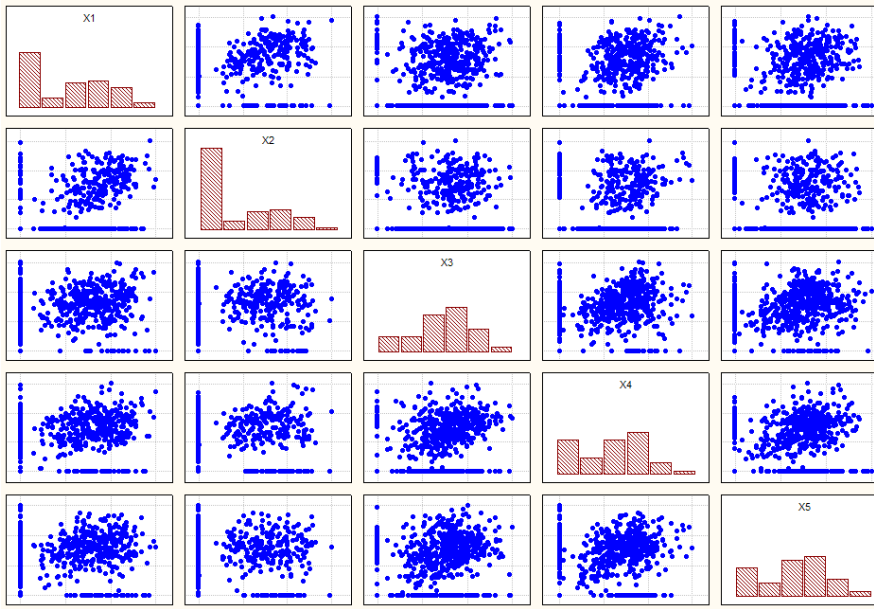


Figure 5. Scatter diagram for transformed features

Spearman coefficient matrix for features **B1–B40** is given in Table 4. The coefficients significant at 0,001 level are marked as red. Most of the features are tightly interconnected. Consequently, there is no reason to use factor analysis of principal components analysis since they are appropriate for correlation matrices with block structure. In the further we apply logistic regression which is invariant under linear transformations so we will use the features themselves, not the principal components or factors.

Table 4. Spearman correlation coefficients for features **B1–B40**

Variable	B3	B4	B5	B6	B7	B8	B9	B10	B11	B12	B13	B14	B15	B16	B17	B18	B19	B20	B21	B22	B23	B24	B25	B26	B27	B28	B29	B30	B31	B32	B33	B34	B35	B36	B37	B38	B39	B40
B1	0,18	0,35	0,23	0,25	0,35	0,13	0,22	0,15	0,28	0,34	0,46	0,22	-0,41	-0,63	-0,49	-0,52	0,30	0,32	0,39	0,16	0,17	-0,10	-0,25	0,33	-0,55	-0,24	-0,05	0,12	-0,50	-0,31	0,29	0,21	-0,44	0,14	0,10	0,05	-0,02	0,44
B2	0,15	0,19	0,13	0,12	0,21	-0,10	0,29	0,19	0,18	0,13	0,30	0,26	-0,28	-0,49	-0,33	-0,34	0,25	0,24	0,28	0,08	0,10	-0,05	-0,20	0,18	-0,42	-0,23	-0,07	0,14	-0,30	-0,29	0,47	0,17	-0,32	0,03	0,02	0,22	0,09	0,32
B3	1,00	0,29	0,26	0,29	0,16	0,31	0,25	0,12	0,06	0,19	0,28	0,16	-0,29	-0,29	-0,30	-0,29	0,20	0,35	0,30	0,23	0,21	0,09	-0,13	-0,12	-0,28	-0,11	-0,06	0,13	-0,33	-0,10	0,07	0,27	-0,37	0,19	0,01	-0,04	0,05	0,30
B4	0,29	1,00	0,27	0,20	0,16	0,13	0,21	0,10	-0,00	0,12	0,24	0,00	-0,22	-0,28	-0,27	-0,26	0,26	0,23	0,25	0,07	0,18	0,03	-0,12	0,03	-0,18	0,02	-0,06	0,17	-0,25	-0,07	0,09	0,21	-0,22	-0,01	0,07	-0,04	0,02	0,23
B5	0,26	0,27	1,00	0,30	0,12	0,15	0,10	0,11	-0,02	0,18	0,24	0,11	-0,15	-0,23	-0,15	-0,25	0,19	0,13	0,25	0,16	0,31	0,06	-0,01	-0,13	-0,13	0,02	-0,06	0,04	-0,22	-0,11	-0,06	0,24	-0,19	0,11	0,07	0,04	-0,01	0,23
B6	0,30	0,29	0,30	1,00	0,40	0,21	0,29	0,14	0,22	0,38	0,40	0,27	-0,25	-0,35	-0,26	-0,20	0,35	0,39	0,44	0,41	0,52	0,24	-0,11	0,03	-0,29	-0,04	0,29	0,22	-0,29	-0,16	0,24	0,38	-0,36	0,34	0,26	0,27	0,28	0,46
B7	0,16	0,16	0,12	0,40	1,00	0,15	0,21	0,20	0,40	0,42	0,44	0,33	-0,34	-0,42	-0,24	-0,06	0,38	0,45	0,44	0,36	0,31	0,14	-0,12	0,14	-0,40	-0,09	0,31	0,20	-0,19	-0,13	0,29	0,28	-0,46	0,35	0,33	0,33	0,22	0,57
B8	0,31	0,13	0,15	0,14	0,15	1,00	0,65	-0,30	0,09	0,26	0,21	0,07	-0,33	0,06	0,07	0,03	0,30	0,35	0,21	0,07	0,11	0,08	-0,26	0,05	0,13	0,15	-0,11	-0,19	0,01	0,10	0,07	0,01	0,03	0,10	0,07	0,01	0,03	0,09
B9	0,25	0,21	0,10	0,29	0,21	0,65	1,00	-0,23	0,23	0,42	0,32	0,25	-0,31	-0,20	-0,08	-0,03	0,45	0,45	0,44	0,39	0,26	0,13	0,11	0,13	-0,32	-0,10	0,23	0,30	-0,14	-0,21	0,42	0,19	-0,46	0,12	0,09	0,25	0,40	0,32
B10	0,12	0,10	0,11	0,14	0,20	-0,30	-0,23	1,00	0,15	-0,01	0,16	0,16	-0,06	-0,21	-0,10	0,01	0,17	0,09	0,18	0,11	0,20	0,20	-0,14	0,14	-0,00	-0,01	0,06	0,19	-0,03	0,18	0,09	0,20	-0,09	0,23	0,12	0,13	0,01	0,23
B11	0,06	-0,00	-0,02	0,22	0,40	0,09	0,23	0,15	1,00	0,43	0,44	0,37	-0,40	-0,35	-0,24	-0,00	0,20	0,31	0,33	0,32	0,21	0,17	-0,15	0,34	-0,47	-0,23	0,48	0,17	-0,20	-0,13	0,41	0,22	-0,37	0,30	0,28	0,22	0,36	0,28
B12	0,19	0,12	0,18	0,38	0,42	0,36	0,42	-0,01	0,43	1,00	0,65	0,15	-0,41	-0,36	-0,30	-0,19	0,30	0,43	0,42	0,36	0,30	0,11	-0,02	0,18	-0,45	-0,08	0,41	0,27	-0,36	-0,21	0,28	0,16	-0,43	0,25	0,28	0,13	0,30	0,41
B13	0,28	0,24	0,24	0,40	0,44	0,21	0,32	0,16	0,44	0,65	1,00	0,08	-0,51	-0,51	-0,41	-0,36	0,44	0,30	0,46	0,19	0,26	0,05	-0,20	0,07	-0,47	-0,24	0,21	0,23	-0,41	-0,28	0,25	0,35	-0,43	0,18	0,29	0,15	0,06	0,52
B14	0,16	0,00	0,11	0,27	0,33	0,07	0,25	0,16	0,37	0,15	0,08	1,00	-0,29	-0,27	-0,28	0,06	0,20	0,38	0,30	0,26	0,13	0,19	-0,09	0,33	-0,12	-0,27	0,43	0,06	-0,23	-0,19	0,15	0,12	-0,35	0,32	0,30	0,36	0,34	0,24
B15	-0,29	-0,22	-0,15	-0,25	-0,34	-0,33	-0,31	-0,06	-0,40	-0,41	-0,51	-0,29	1,00	0,34	0,44	0,35	-0,20	-0,38	-0,37	-0,24	-0,27	0,03	0,25	-0,16	0,63	0,31	-0,22	-0,19	0,46	0,34	-0,29	-0,25	0,48	-0,22	-0,18	-0,06	-0,09	-0,42
B16	-0,29	-0,28	-0,23	-0,35	-0,42	-0,06	-0,20	-0,21	-0,35	-0,36	-0,61	-0,27	0,34	1,00	0,42	0,45	-0,39	-0,33	-0,43	-0,16	-0,19	0,05	0,32	-0,20	0,62	0,27	-0,01	0,16	0,41	0,34	-0,34	-0,34	0,43	-0,18	-0,19	-0,26	-0,08	-0,54
B17	-0,30	-0,27	-0,15	-0,26	-0,24	0,03	-0,08	-0,10	-0,24	-0,30	-0,41	-0,28	0,44	0,42	1,00	0,49	-0,26	-0,25	-0,28	-0,22	-0,20	0,12	0,31	-0,16	0,48	0,28	-0,04	0,16	0,88	0,33	-0,24	0,18	0,39	-0,13	-0,19	-0,03	-0,03	-0,36
B18	-0,29	-0,26	-0,25	-0,20	-0,06	-0,07	-0,03	0,01	-0,00	-0,19	-0,36	0,06	0,35	0,45	0,49	1,00	-0,12	-0,11	-0,09	0,00	-0,09	0,42	0,34	0,25	0,38	0,29	0,29	-0,02	0,60	0,38	-0,01	-0,09	0,24	0,22	0,24	0,28	0,30	-0,15
B19	0,20	0,26	0,19	0,35	0,38	0,03	0,45	0,17	0,20	0,30	0,44	0,20	-0,20	-0,39	-0,26	-0,12	1,00	0,29	0,39	0,25	0,25	0,06	-0,03	0,07	-0,27	-0,06	0,16	0,23	-0,26	-0,14	0,30	0,38	-0,40	0,17	0,11	0,43	0,48	0,45
B20	0,35	0,23	0,13	0,39	0,45	0,42	0,45	0,09	0,31	0,43	0,30	0,38	-0,33	-0,25	-0,11	0,29	1,00	0,45	0,47	0,30	0,14	-0,06	0,33	0,41	-0,20	0,36	0,31	0,29	-0,25	0,36	0,28	-0,55	0,36	0,19	0,23	0,18	0,45
B21	0,30	0,25	0,25	0,44	0,44	0,30	0,44	0,18	0,33	0,42	0,46	0,30	-0,37	-0,43	-0,28	-0,09	0,39	0,45	1,00	0,49	0,33	0,15	-0,11	0,22	-0,42	-0,21	0,30	0,28	-0,29	-0,16	0,37	0,31	-0,47	0,23	0,10	0,27	0,24	0,44
B22	0,23	0,07	0,16	0,41	0,39	0,35	0,39	0,11	0,32	0,36	0,19	0,26	-0,24	-0,19	-0,22	0,00	0,25	0,47	0,49	1,00	0,35	0,27	-0,00	0,16	-0,28	-0,09	0,41	0,29	-0,24	-0,13	0,26	0,29	-0,40	0,39	0,19	0,31	0,45	0,34
B23	0,21	0,18	0,21	0,52	0,31	0,21	0,26	0,20	0,21	0,30	0,26	0,13	-0,27	-0,19	-0,20	-0,09	0,25	0,30	0,33	0,35	1,00	0,28	-0,06	-0,07	-0,19	0,01	0,36	0,24	-0,23	-0,15	0,16	0,28	-0,30	0,33	0,19	0,12	0,25	0,31
B24	0,09	0,03	0,06	0,24	0,14	0,07	0,13	0,20	0,17	0,11	0,05	0,19	0,03	0,05	0,12	0,42	0,06	0,14	0,15	0,27	0,28	1,00	0,22	0,08	0,01	0,13	0,48	0,21	0,13	0,15	0,26	0,11	-0,14	0,35	0,28	0,22	0,37	0,10
B25	-0,13	-0,12	-0,01	-0,11	-0,12	0,11	0,11	-0,14	-0,15	-0,02	-0,20	-0,09	0,25	0,32	0,31	0,34	-0,03	-0,06	-0,11	0,00	-0,06	0,22	1,00	-0,08	0,17	0,21	0,01	-0,06	0,29	0,20	-0,16	-0,19	0,06	-0,03	0,05	0,04	0,06	-0,20
B26	-0,12	0,03	-0,13	-0,03	0,14	-0,08	0,13	0,14	0,34	0,18	0,07	0,33	-0,16	-0,20	-0,16	0,25	0,07	0,33	0,22	0,16	-0,07	0,08	-0,08	1,00	-0,36	-0,15	0,46	0,35	-0,13	-0,01	0,48	-0,00	0,29	-0,35	0,07	0,22	0,24	0,20
B27	-0,28	-0,18	-0,13	-0,29	-0,40	-0,26	-0,32	-0,00	-0,47	-0,45	-0,47	-0,32	0,63	0,62	0,48	0,38	-0,27	-0,41	-0,42	-0,28	-0,19	0,01	0,17	-0,36	1,00	0,36	-0,13	-0,20	0,51	0,48	-0,41	-0,23	0,70	-0,22	-0,21	-0,24	-0,18	-0,45
B28	-0,11	0,02	0,02	-0,04	-0,09	-0,05	-0,10	-0,01	-0,23	-0,08	-0,24	-0,27	0,31	0,27	0,28	0,29	-0,06	-0,20	-0,21	-0,09	0,01	0,13	0,21	-0,15	0,36	1,00	-0,16	-0,03	0,30	0,34	-0,28	-0,14	0,23	0,05	0,16	0,06	0,10	-0,19
B29	-0,06	-0,06	-0,06	0,29	0,31	0,13	0,23	0,06	0,48	0,41	0,21	0,43	-0,22	-0,01	-0,04	0,29	0,16	0,36	0,30	0,41	0,36	0,48	0,01	0,46	-0,13	-0,16	1,00	0,48	-0,04	0,08	0,48	0,32	-0,09	0,60	0,83	0,33	0,59	0,20
B30	0,13	0,17	0,04	0,22	0,20	0,16	0,30	0,19	0,17	0,27	0,23	0,06	-0,19	-0,16	-0,16	-0,02	0,23	0,31	0,28	0,29	0,24	0,21	-0,06	0,35	-0,20	-0,03	0,48	1,00	-0,16	0,05	0,31	0,17	-0,19	0,18	0,17	0,10	0,28	0,22
B31	-0,33	-0,25	-0,22	-0,29	-0,19	-0,15	-0,14	-0,03	-0,20	-0,36	-0,41	-0,23	0,46	0,41	0,88	0,60	-0,26	-0,29	-0,29	-0,24	-0,23	0,13	0,29	-0,13	0,51	0,30	-0,04	-0,16	1,00	0,40	-0,20	-0,21	0,44	-0,09	-0,21	-0,00	-0,06	-0,34
B32	-0,10	-0,07	-0,11	-0,16	-0,13	-0,18	-0,21	0,18	-0,13	-0,21	-0,28	-0,19	0,34	0,34	0,33	0,38	-0,14	-0,25	-0,16	-0,13	-0,15	0,15	0,20	-0,01	0,46	0,34	0,08	0,05	0,40	1,00	-0,18	-0,09	0,34	0,03	0,04	-0,04	-0,08	-0,25
B33	0,07	0,09	-0,06	0,24	0,29	-0,01	0,42	0,09	0,41	0,28	0,25	0,35	-0,29	-0,34	-0,24	-0,01	0,30	0,36	0,37	0,26	0,16	0,26	-0,16	0,48	-0,41	-0,28	0,48	0,31	-0,20	-0,18	1,00	0,18	-0,34	0,36	0,23	0,29	0,34	0,33
B34	0,27	0,21	0,24	0,38	0,28	0,19	0,19	0,20	0,22	0,19	0,35	0,12	-0,25	-0,34	-0,18	-0,09	0,38	0,28	0,31	0,29	0,28	0,11	-0,18	-0,00	-0,23	-0,14	0,32	0,17	-0,21	-0,09	0,18	1,00	-0,27	0,39	0,21	0,32	0,27	0,35
B35	-0,37	-0,22	-0,19	-0,36	-0,46	-0,38	-0,46	-0,09	-0,37	-0,43	-0,43	-0,35	0,48	0,43	0,39	0,24	-0,40	-0,55	-0,47	-0,40	-0,30	-0,14	0,06	-0,20	0,70	0,22	-0,09	-0,19	0,44	0,34	-0,34	-0,27	1,00	-0,29	-0,23	-0,29	-0,23	-0,48
B36	0,19	-0,01	0,11	0,34	0,35	0,07	0,12	0,23	0,30	0,25	0,18	0,32	-0,22	-0,18	-0,13	0,22	0,17	0,36	0,23	0,39	0,33	0,35	-0,03	0,35	-0,22	0,05	0,60	0,18	-0,09	0,03	0,36	0,39	-0,29	1,00	0,54	0,36	0,35	0,32
B37	0,01	0,07	0,07	0,26	0,33	0,01	0,09	0,12	0,28	0,29	0,29	0,30	-0,18	-0,19	-0,19	0,24	0,11	0,19	0,10	0,19	0,19	0,28	0,05	0,07	-0,21	0,16	0,63	0,17	-0,21	0,04	0,23	0,21	-0,23	0,54	1,00	0,29	0,30	0,31
B38	-0,04	-0,04	0,04	0,27	0,33	-0,13	0,25	0,13	0,22	0,13	0,15	0,36	-0,06	-0,26	-0,03	0,28	0,43	0,23	0,27	0,31	0,12	0,22	-0,04	0,22	-0,24	-0,06	0,33	0,10	-0,00	-0,04	0,29	0,32	-0,29	0,36	0,29	1,00	0,45	0,30
B39	0,06	0,02	-0,01	0,28	0,22	0,09	0,40	0,01	0,36	0,30	0,06	0,34	-0,09	-0,08	-0,03	0,30	0,46	0,18	0,24	0,45	0,25	0,37	0,06	0,24	-0,18	0,10	0,59	0,28	-0,06	-0,08	0,34	0,27	-0,23	0,35	0,30	0,45	1,00	0,12
B40	0,30	0,23	0,23	0,46	0,57	0,21	0,32	0,23	0,28	0,41	0,52	0,24	-0,42	-0,54	-0,36	-0,15	0,45	0,45	0,44	0,34	0,31	0,10	-0,20	0,20	-0,45	-0,19	0,20	0,22	-0,34	-0,25	0,33	0,35	-0,48	0,32	0,31	0,30	0,12	1,00

Class prediction

To achieve better robustness we removed 86 (13,9%) texts of gold standard, which we not assigned to any cluster and consequently were considered as noise. For every class among **C1–C9** we constructed a single logistic model (Bishop, 2006) separating it from the other classes. For every class we used its corpus frequency to set classification threshold. The values of thresholds, sensitivity and specificity are given in Table 5.

Table 5. Threshold, sensitivity and specificity for classes **C1–C9**

	C1	C2	C3	C4	C5	C6	C7	C8	C9
Threshold τ	0,11	0,13	0,14	0,13	0,08	0,14	0,09	0,07	0,13
Sensitivity	81%	81%	92%	94%	94%	85%	81%	94%	85%
Specificity	86%	88%	93%	97%	90%	88%	85%	94%	83%

For multiclass classification we used one-versus-all (OVA) approach: a point was assigned to the class whose separation hyperplane was on the largest signed distance. Distance was measured according to the formula below; negative value means that logistic model does not assign this point to a class under consideration.

$$d = \ln \frac{\left(\frac{1}{\tau} - 1\right) / \left(\frac{1}{\rho} - 1\right)}{\sqrt{b_1^2 + \ldots + b_1^2}},\ b_1, \ldots, b_9 \text{ — } \textit{feature weights in regression model.}$$

We obtained the following table of cross-classification:

Table 6. Table of cross-classification for classes **C1–C9**

Class	C1	C2	C3	C4	C5	C6	C7	C8	C9	Total
C1	37	9	3	0	3	1	1	1	1	**56**
C2	9	41	0	0	2	3	2	1	9	**67**
C3	6	0	59	0	0	0	7	0	1	**73**
C4	0	0	1	64	0	3	0	1	0	**69**
C5	0	0	0	0	34	0	4	2	1	**41**
C6	1	0	1	4	3	57	9	0	0	**75**
C7	4	3	4	1	0	1	30	1	3	**47**
C8	1	0	0	1	1	0	0	30	2	**35**
C9	4	6	1	0	2	1	2	3	50	**69**
Accuracy	66%	61%	81%	93%	83%	76%	64%	86%	72%	**76%**
	78%		81%	93%	83%	80%		86%	72%	**81%**

The pairs **C1, C2** and **C6, C7** are the most difficult to separate. If we join the elements of each pair together, accuracy grows from 76% to 81%. Unification of **C6** and C7 is justified from linguistical point of view also: the class **C7** contains texts with principal dimension **A16** (at first, encyclopedic), whereas **C6** comprises documents

with principal FTDs **A14** (scientific or technical) and **A15** (texts for specialists). For example, scientific encyclopedic articles are on the border; therefore, these classes cannot be reliably distinguished not only automatically but by human annotator also. Classification accuracy for all classes lies in the diapason **72–93%**. Thus, logistic regression model permits to perform automatic genre classification with high accuracy at least for the texts of gold standard.

Predicting functional text dimensions

Almost surely, new clusters will emerge for a corpus of other size or origin. Therefore an automatic procedure should be designed to detect new clusters in data and assign texts to these clusters, as well as to separate dense regions from ubiquitous noise. To address these tasks the values of the very FTDs should be known. Again we used logistic regression, but in its weighted variant: when predicting a feature, the texts with rate 2 for this feature had weight 2. Inversed backward procedure was used to increase stability. As earlier, we used the frequency of a feature to determine classification threshold. The values of threshold, sensitivity and specificity for 16 dimensions are given in Table 7. The test values of performance measures were calculated using 70%/30% train-test split.

Table 7. Accuracy of automatic FTD prediction

	A1	A3	A4	A5	A6	A7	A8	A9	A11	A12	A13	A14	A15	A16	A17	A18
Threshold	0,33	0,23	0,09	0,14	0,23	0,11	0,25	0,22	0,31	0,20	0,19	0,18	0,20	0,61	0,24	0,08
Sensitivity full	76%	88%	92%	79%	82%	94%	86%	96%	83%	89%	74%	86%	71%	76%	72%	91%
Specificity full	78%	91%	96%	80%	83%	93%	86%	97%	86%	90%	80%	89%	75%	77%	83%	96%
Sensitivity test	79%	94%	92%	76%	81%	91%	80%	98%	79%	90%	72%	91%	73%	72%	71%	89%
Specificity test	67%	76%	100%	81%	72%	95%	73%	83%	90%	80%	64%	71%	75%	66%	88%	100%

Classification accuracy is quite high. Dimensions **A1**, **A13**, **A16** were predicted a bit worse than others on the test sample. Average specificity on the full data and test sample was 83% and average sensitivity was 86% and 80%, respectively. To make the model even more reliable we selected only the features which are significant at p-level=0,05. Since features are standartized to have the same mean and variance, the absolute values of weights reflect the importance of the corresponding features for a predicted dimension. Logistic model also predicts class probabilities. For example, the probabilities below confidently attribute text to the class **C4** (**A9**+**A16**).

A1	A3	A4	A5	A6	A7	A8	A9	A11	A12	A13	A14	A15	A16	A17	A18
0,024	0,001	0,000	0,011	0,021	0,008	0,011	0,962	0,003	0,535	0,005	0,000	0,247	0,728	0,003	0,000

FTD probability scores being known, the easiest way to classify text is the following: multiply all the probabilities by 2 and assign text to the closest prototype in the sense of Euclidean distance. When this distance exceeds 2, a text is considered as noise. We refer to such classification mechanism as crude classification. The results of crude classification of FTDs for gold standard are given below. They are quite more decent than before, which justifies the term “crude classification”.

Table 8. Results of crude classification

	Noise	C1	C2	C3	C4	C5	C6	C7	C8	C9	Total
Accuracy	43%	45%	31%	64%	94%	66%	57%	55%	80%	57%	**58%**
	43%	48%		64%	94%	66%	80%		80%	57%	**65%**

Such method allows us to detect 43% of noise. Investigating its structure more thoroughly, we see that most of the noise texts are rare combinations of 2 or 3 functional dimensions. Such combinations may potentially become new cluster centers for a more representative corpus.

Rank	0	1	2	3	4	5	6
Count	3	5	44	23	8	2	1
Percentage	3%	6%	51%	27%	9%	2%	1%

Application to a large corpus

The ultimate goal of our research is to prepare the corpus for automatic genre annotation. The results of automatic classification are rather optimistic in this sense, showing that FTD values could in principle be predicted reliably. However, we need to check whether the model learnt on the gold standard would be correct on another corpus. For this task we should manually annotate another corpus and compare the values of FTDs with those predicted by the logistic models. Since this comparison is time and labour-consuming, we just explore the combinations of features more frequent on a new corpus.

1) We took 1,000,000 LJ-posts from the current version of GICR and applied the logistic models learnt on the gold standard. This yields 16 probability scores for every text. For the sake of simplicity we performed binary classification without intermediate values.
2) We binarized the scores for every feature using the same threshold 0,75. Thus every text is mapped to a 16-dimensional binary vector, which naturally corresponds to a set of FTDs of this text. The vectors which occur more than 1000 times in the corpora were considered as the prototypes.
3) We attached the texts to the closest prototype by the Euclidean distance between the probability scores of the text and the prototype. The texts with no prototype within the distance 1 were not attached to any cluster. Here a cluster is a set of texts with the same prototype.

We detected 120 prototypes with ranks varying from 0 to 5 (rank is the number of positive FTDs). For 208,533 of 1,000,000 texts no prototype was found. Using the "knee method" of (Salvador, 2004) we discovered the 9 most frequent clusters which are the prototypes for 266,635 texts.

Table 9. The most frequent combinations of FTD values after crude classification

Feature combination	Texts number
A8	49,738
A3+A6+A11	35,506
A16	34,578
A3+A6+A11+A17	31,214
A8+A16	30,219
A3+A5+A6+A11+A17	23,036
A1+A3+A6+A11	20,917
A12+A16	20,874
A3+A4+A6+A11	20,553

Some of the combinations (**A8**, **A16**, **A12+A16** etc.) occur already in the gold standard; while the combinations of features **A3**, **A6**, **A11** increased their frequency. This property is the most noticeable for FTD A17 "evaluation". This results are quite expectable since argumentative (**A1**), informal (**A6**) and evaluative (**A17**) texts are usual in blogosphere. However, additional verification of assigned rates is necessary.

Conclusions

While straightforward classification of a large Web corpus into genres is problematic because of the disagreement between the human annotations, we managed to provide an overview of the most frequent genre options in the corpus. Overall, our research leads to the following conclusions:

1) There exist well-formed dense clusters in the FTD space.
2) Language features allow prediction of the FTD values with high accuracy (about 75%).
3) We can detect very similar clusters when the model learnt on the gold standard are applied to a new corpus.

The achieved results are rather promising, but there is still a long way to go before achieving reliable genre annotation of large real-world corpora. Our plan is to replace traditional genres with FTD clusters, and the present research demonstrates, that such clusters can be detected automatically. However, the list of pseudogenres observed in the "gold standard" is obviously incomplete, so that we need to uncover the cluster structure of bigger corpora in future research.

In future genre annotation experiments we will match the texts in the corpus with the label of the closest cluster in the FTD space. For some texts the classifier can be confident, assigning scores about 1.0 to the principal dimensions and near zero scores for other FTDs. Hence we can assign a cluster label (therefore, genre label) for such texts reliably. For other texts the classifier can be less confident, and the best strategy would be to refuse attributing any genre label or to give a list of nearby clusters. Therefore, we plan to continue our research in the following directions:

1. To collect a near-exhaustive list of possible clusters in the FTD space using a bigger corpus.
2. Check the accuracy of automatic detection of these clusters
3. Consider the centers of such clusters as prototypes
4. Assign the documents of the corpora to their closest prototype provided the prototype is indeed close in the FTD space

Though this procedure definitively would not allow to detect genre for arbitrary Web text, we plan to assign labels to a large fraction of the corpus with sufficient confidence. Even incomplete annotation gives a possibility to study e.g genre-dependent linguistic parameters and the linguistic correlates of genre labels.

References

1. *Biber, D.* Variation across speech and writing. Cambridge: CUP, 1988
2. *Biber D., Connor, U. and Upton, T.* Discourse on the move: using corpus analysis to describe discourse structure. Amsterdam—Philadelphia, 2007.
3. *Bishop C.* Pattern recognition and machine learning. Springer, New York, 2006.
4. *Egbert J., Biber D.* Developing a user-based method of register classification // Proc. 8th Web as Corpus Workshop. Lancaster, 2013. July.
5. *Forsyth R., Sharoff S.* (2014). Document dissimilarity within and across languages: a benchmarking study // Literary and Linguistic Computing. — 2014. — 29(1):6–22.
6. *Kilgariff A.* The Web as corpus // Proc. of corpus linguistics 2001. Lancaster, 2001.
7. *Kutner M., Nachtsheim C., Neter J., and Li W.* (2004). Applied Linear Statistical Models, McGraw-Hill/Irwin, Homewood, IL, 2004.
8. *Salvador S., Chan P.* // Proc. of 16 IEEE International Conference on Tools with Artificial Intelligence (2004). P. 576–584.
9. *Sharoff S.* In the garden and in the jungle: Comparing genres in the BNC and the Internet // Genres on the Web: Computational Models and Empirical Studies / Ed. by Alexander Mehler, Serge Sharoff, Marina Santini. Berlin/New York : Springer, 2010. P. 149–166.
10. *Sorokin A., Katinskaya A., Sharoff S.* Associating symptoms with syndromes: Reliable genre annotation for a large Russian webcorpus // Proc. Dialogue, Russian International Conference on Computational Linguistics. Bekasovo, 2014.
11. *Беликов В. И., Селегей В. П., Шаров С. А.* Пролегомены к проекту Генерального интернет-корпуса русского языка. // Труды конференции Диалог 2012.

Appendix 1. Functional Text Dimensions

Code	Label	Question to be answered
A1	argum	To what extent does the text argue to persuade the reader to support (or renounce) an opinion or a point of view? ('Strongly', if argumentation is obvious)
A3	emotive	To what extent is the text concerned with expressing feelings or emotions? ('None' for neutral explanations, descriptions and/or reportage.)
A4	fictive	To what extent is the text's content fictional? ('None' if you judge it to be factual/informative.)
A5	flippant	To what extent is the text light-hearted, i.e. aimed mainly at amusing or entertaining the reader? ('None' if it appears earnest or serious; even when it tries to keep the reader interested and involved)
A6	informal	To what extent is the text's content written in an informal style? (as opposed to the «standard» or «prestige» variety of language)
A7	instruct	To what extent does the text aim at teaching the reader how to do something? (For example, a tutorial or an FAQ)
A8	hardnews	To what extent does the text appear to be an informative report of recent events? (Recent at the time of writing. Announcements of future events can be considered hardnews too. 'None' if a news article only analyses information from other sources).
A9	legal	To what extent does the text lay down a contract or specify a set of regulations? (This includes copyright notices.)
A11	personal	To what extent does the text report from a first-person point of view? (For example, a diary-like blog entry.)
A12	compuff	To what extent does the text promote a product or service?
A13	ideopuff	To what extent is the text intended to promote a political movement, party, religious faith or other non-commercial cause?
A14	scitech	To what extent would you consider the text as representing research? (It does not have to be a research paper. For example, 'Strongly' or 'Partly' if a newswire text has scientific contents.)
A15	specialist	To what extent does the text require background knowledge or access to a reference source of a specialised subject area in order to be comprehensible? (such as wouldn't be expected of the so-called "general reader")
A16	info	To what extent does the text provide information to define a topic? (For example, encyclopedic articles or text books).
A17	eval	To what extent does the text evaluate a specific entity by endorsing or criticising it? (For example, by providing a product review).
A18	dialogue	To what extent does the text contain active interaction between several participants? (For example, forums or scripted dialogues).
A19	poetic	To what extent does the author of the text pay attention to its aesthetic appearance? ('Strongly' for poetry, language experiments, uses of language for art purposes).

Appendix 2. Linguistic features for automatic classification

B1	first_person_pronoun
B2	second_person_pronoun
B3	third_person_pronoun
B4	reflexive_pronoun
B5	adjective_pronoun
B6	nom_pronoun
B7	indefinite_pron
B8	past_tense
B9	perf_aspect
B10	present_tense
B11	place_adverb
B12	time_adverb
B13	total_adverb
B14	wh_questions
B15	nominalization
B16	nouns
B17	passive
B18	by_passive
B19	infinitive
B20	speech_verb

B21	mental_verb
B22	that_compl
B23	wh_relative
B24	pied_piping
B25	total_PP
B26	exclamation
B27	word_length
B28	type_token_ratio
B29	sentence_length
B30	verbal_adverb
B31	passive_participial_clauses
B32	active_participial_clauses
B33	imperative_mood
B34	predicative_adjectives
B35	attributive_adjective
B36	causative_subordinate
B37	concessive_subordinate
B38	conditional_subordinate
B39	purpose_subordinate
B40	Negation

ОПЫТ СОЗДАНИЯ МЕЛОДИЧЕСКИХ ПОРТРЕТОВ СЛОЖНЫХ ПОВЕСТВОВАТЕЛЬНЫХ ПРЕДЛОЖЕНИЙ РУССКОЙ РЕЧИ

Лобанов Б. М. (Lobanov@newman.bas-net.by)

Объединённый институт проблем информатики
НАН Беларуси, Минск, Беларусь

Ключевые слова: интонационные конструкции, мелодический портрет, синтез и анализ интонации, русская интонации, русский как иностранный

AN EXPERIENCE OF CREATING MELODIC PORTRAITS OF COMPLEX DECLARATIVE SENTENCES OF RUSSIAN

Lobanov B. M. (Lobanov@newman.bas-net.by)

United Institute of Informatics Problems of NAS Belarus,
Minsk, Belarus

We proceed from the model of intonation patterns (IP) by Elena Bryzgunova, widely used in teaching Russian speech intonation. Bryzgunova distinguishes seven major Russian intonation patterns, named IP 1 to IP 7, of which only IP 1 is clearly used in declarative sentences to mark their completeness. The remaining six IPs are implemented for interrogative (IP2 — IP 4) or exclamatory (IP5 — IP7) types of sentences. Obviously, declarative sentences are overwhelming in professional and literary texts, particularly in professionally voiced texts of various genres (audio books). Most of them are not simple sentences and often consist of a mixture of complex and compound sentences.

The present study continues the author's paper "Universal melodic intonation portraits of Russian speech" presented to Dialogue 2014 conference, which introduced the concept of Universal Melodic Portrait (UMP). The present paper experimentally studies the intonation features of declarative sentences. It describes the results of auditory analysis and IP interpretation for declarative sentences of varying degrees of complexity, voiced by 3 speakers, and provides experimental representations of their intonation structures in the form of a sequence of Universal Melodic Portraits (UMP).

The paper is organized as follows. Section 1 describes the experimental procedure, including the characteristics of selected text and audio material, the listening method and the method of constructing a sequence

of UMP's audio recordings. Section 2 presents the experimental results: the graphical representation of an experimental sequence of universal melodic portraits of analyzed audio recordings. Section 3 offers an interpretation of the results.

Keywords: intonation patterns, melodic portrait, synthesis and analysis of intonation, Russian intonation, Russian as the second language

Введение

В 1960-х гг. Е. А. Брызгунова предложила описание интонации русского языка [Брызгунова, 1968] с использованием понятия *интонационной конструкции* (ИК), которое вошло в академическое издание русской грамматики и стало повсеместно использоваться методических пособиях по обучению русского языка как иностранного (РКИ) [*Одинцова*, 2011].

В предыдущей работе автора [Лобанов, 2014] дано обоснование представления семи интонационных конструкций Брызгуновой в виде набора Универсальных мелодических портретов (УМП). В процессе публикации этой работы одним из рецензентов высказано следующее сомнение в широте применимости предлагаемых УМП:

> *Мне очень близка мысль автора доклада, что предложенные Е. А. Брызгуновой интонационные конструкции, ставшие уже хрестоматийными и прочно занявшие свое место в методике преподавания РКИ, не могут удовлетворить «разработчиков компьютерных моделей анализа и синтеза интонационных характеристик речи».*
>
> *От себя добавлю: и исследователей живой устной речи, в которой таких эталонных ИК практически не существует. То, что хорошо для начального знакомства с интонационной системой русского языка, абсолютно недостаточно во многих других, прежде всего исследовательских, целях.*

Действительно, Е. А. Брызгунова выделяет семь основных интонационных конструкций русского языка: ИК1 — ИК7. При этом только ИК1 явно используется в повествовательных предложениях для выражения его завершенности. Примеры реализации остальных ИК приводятся для вопросительных (ИК2 — ИК4) или восклицательных (ИК5 — ИК7) предложений. В то же время очевидно, что повествовательные предложения составляют подавляющее большинство в дикторской речи, в частности, в профессионально озвучиваемых текстах различного жанра. При этом зачастую большинство из них не являются простыми предложениями и состоят из совокупности сложносочинённых и сложноподчинённых распространённых предложений.

В настоящей работе дано экспериментальное представление интонационной структуры набора повествовательных предложений различной степени сложности в виде описанной ранее последовательности УМП. Приводятся результаты аудитивного анализа и ИК-интерпретации построенной последовательности УМП для семи повествовательных предложений различной степени сложности, озвученных тремя профессиональными дикторами.

Настоящая работа построена следующим образом.

В первом разделе описывается выбранный текстовый и аудиоматериал, методика аудирования и способ построения последовательности УМП исследуемых аудиозаписей.

Во втором разделе приводятся в графическом виде результаты построения экспериментальных УМП и данные об ИК разметки проанализированных текстов.

Третий раздел посвящен анализу и интерпретации полученных результатов.

1. Методика эксперимента

Для проведения эксперимента выбраны первые 6 повествовательных предложений различной степени сложности из повести А. П. Чехова «Драма на охоте» и одно сложное повествовательное предложение из его же повести «Дама с собачкой». Обе повести имели аудиоверсии в исполнении профессиональных дикторов. Аудиокнига повести «Драма на охоте» исполнена диктором Александром Балакиревым. Анализировалась интонация чтения следующих 6-ти предложений:

(1) *В один из апрельских полудней тысяча восемьсот восьмидесятого года в мой кабинет вошел сторож Андрей и таинственно доложил мне, что в редакцию явился какой-то господин и убедительно просит свидания с редактором.*

(2) *Начинающие писатели и вообще люди, не посвященные в редакционные тайны, приходящие при слове редакция в священный трепет, заставляют ждать себя немалое время.*

(3) *Они, после редакторского «проси», долго кашляют, долго сморкаются, медленно отворяют дверь, еще медленнее входят и этим отнимают немало времени.*

(4) *Не успела за Андреем затвориться дверь, как я увидел в своем кабинете высокого широкоплечего мужчину, державшего в одной руке бумажный сверток, а в другой — фуражку с кокардой.*

(5) *Одет он со вкусом и по последней моде в новенький, недавно сшитый триковый костюм.*

(6) *Лицо розовое, руки велики, грудь широкая, мускулистая, волосы густы, как у здорового мальчика.*

Для анализа интонации сложного повествовательного предложения из повести «Дама с собачкой» использованы 2 различные версии аудиокниг: в исполнении дикторов Станислава Концевича и Владислава Ветрова. Анализировалась интонация чтения следующего предложения:

(7) *Опыт многократный, в самом деле горький опыт, научил его давно, что всякое сближение, которое вначале так приятно разнообразит жизнь и представляется милым и легким приключением, у порядочных людей, особенно у москвичей, тяжелых на подъем, нерешительных, неизбежно вырастает в целую задачу, сложную чрезвычайно, и положение в конце концов становится тягостным.*

Аудиоверсии чтения 6-ти предложений Александром Балакиревым и 2 варианта чтения 7-го предложения Станиславом Концевичем и Владиславом Ветровым предлагались для анализа последовательно двум независимым аудиторам, имеющим многолетний опыт экспериментально-фонетических исследований. Первому аудитору предъявлялись исходные тексты предложений и их неразмеченные аудиоверсии. В задачу аудитора входило разбиение предъявленных аудиоверсий предложений на просодические синтагмы с указанием акцентных единиц (АЕ), входящих в состав каждой синтагмы. Результат аудирования представлялся в следующем виде:

[В оди+н] [из апре+льских] [полу+дней] #
[ты+сяча] [восемьсо+т] [восьмидесятого го+да]#
[в мо+й] [кабине+т] #
[воше+л] [сторож Андре+й] #
[и таи+нственно] [доложи+л мне], #
[что в реда+кцию] #
[яви+лся] [какой-то господи+н]#
[и убеди+тельно] [про+сит] #
[свида+ния] [с реда+ктором].

Здесь знаком # обозначены межсинтагменные границы (не всегда сопровождаемые паузой), квадратными скобками выделены АЕ, входящие в состав каждой из синтагм, знаком + обозначена позиция ядерной гласной в АЕ.

Второму аудитору предъявлялись тексты предложений с указанием границ синтагм (но без границ АЕ) и их неразмеченные на синтагмы аудиоверсии. В задачу аудитора входило указать, к какой из интонационных конструкций (ИК1 — ИК7) принадлежат входящие в синтагму слова. Результат аудирования представлялся в следующем виде:

В один^{6} *из апрельских полудней #*
тысяча восемьсот восьмидесятого ***года***^{3}*#*
в мой^{6} *кабинет #*
вошел сторож ***Андрей***^{6} *#*

и таинственно ***доложил***[6] *мне, #*
что в ***редакцию***[3] *#*
явился какой-то ***господин***[2] *#*
и ***убедительно***[6] *просит #*
свидания с ***редактором***[1].

Здесь цифровые индексы обозначают номер ИК по Е. А. Брызгуновой, приписываемый выделенному жирным шрифтом слову.

Результаты разбиения предложений на синтагмы и маркировки АЕ, выполненные 1-м аудитором, использовались для построения последовательности УМП каждой синтагмы. Методика построения УМП (с использованием специальных программных средств — «ФОНОКЛОНАТОР» и «ИНТОКЛОНАТОР») достаточно подробно описана в [Лобанов, 2014]. При этом с использованием ФОНОКЛОНАТОРа автоматизируется процесс фонемной сегментации и маркировки каждой АЕ на предъядро, ядро и заядро, а с использованием ИНТОКЛОНАТОРа автоматизируется процесс формирования УМП АЕ.

2. Результаты эксперимента

Основной задачей является построение для выбранного набора повествовательных предложений экспериментальной последовательности {УМП AEi} с их дальнейшим сопоставлением с результатами субъективного слухового отождествления с одной из 7-ми ИК. Полученные результаты представлены на рисунках 1–6 для первых 6-ти повествовательных предложений из «Драмы на охоте» в исполнении диктора Александра Балакирева, а на рисунках 7, 8 — для предложения из повести «Дама с собачкой» в исполнении, соответственно, диктора Станислава Концевича и Владислава Ветрова.

На каждом из рисунков текст предложения записан в двух видах:

- в верхних строках помещается результат разбиения предложений 1-м аудитором на просодические синтагмы с указанием АЕ, входящих в состав каждой синтагмы, а также позиции ядра АЕ;
- в нижних строках помещается результат слухового отождествления 2-м аудитором входящих в синтагму слов с одной из 7-ми ИК.

Над текстом представлена графически последовательность УМП, описывающих каждую АЕ. При этом УМП представлены в нормированных координатах «Частота — Время». Интервалам на оси абсцисс соответствуют: [0–1/3] — предъядро, [1/3–2/3] — ядро, [2/3–1] — заядро, а нормированная частота основного тона на оси абсцисс изменяется от 0 до 1 и рассчитывается по формуле:

$$F_0^N = (F_0 - F_{0\,min}) / (F_{0\,max} - F_{0\,min}).$$

Значения $F_{0\,min}$ и $F_{0\,max}$ для исследуемых дикторов приведены в таблице 1.

Таблица 1

	А. Балакирев	**С. Концевич**	**В. Ветров**
$F_{0\,\min}$ [Hz]	65	60	55
$F_{0\,\max}$ [Hz]	240	300	180

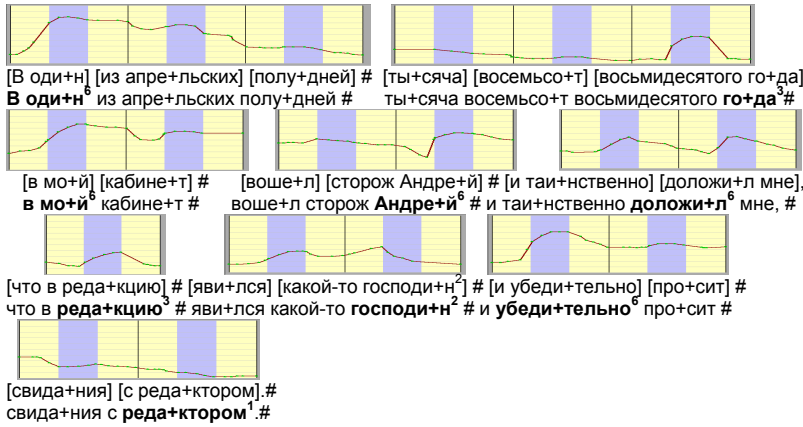


Рис. 1. Последовательность УМП предложения 1 (диктор А. Балакирев)

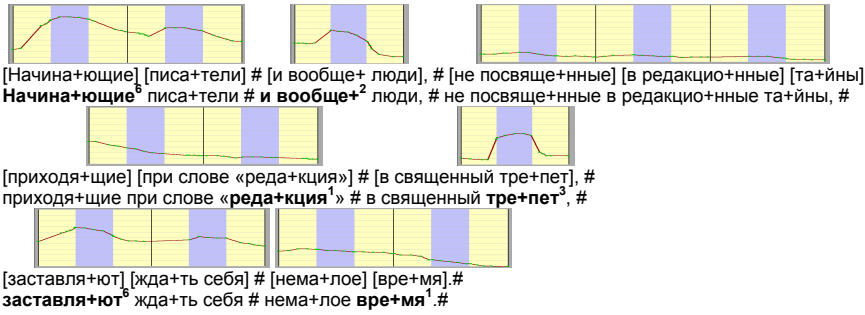


Рис. 2. Последовательность УМП предложения 2 (диктор А. Балакирев)

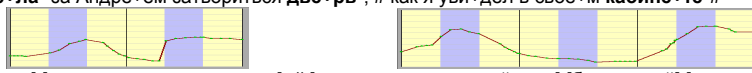

[Они+, после] # [реда+кторского] # [«проси+»], # [долго ка+шляют], # [долго сморка+ются], #
Они+, после # реда+кторского # «**проси+**[3]», # долго **ка+шляют**[3], # долго **сморка+ются**[3], #


[ме+длено] [отворяют две+рь], # [еще+ медленнее] [вхо+дят] #
ме+длено отворяют **две+рь**[3], # **еще+**[2] медленнее **вхо+дят**[6] #


[и э+тим] [отнима+ют] # [нема+ло] [вре+мени].#
и э+тим отнима+ют # нема+ло **вре+мени**[1].#

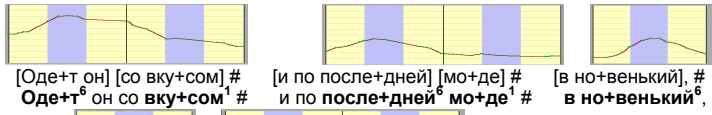
Рис. 3. Последовательность УМП предложения 3 (диктор А. Балакирев)


[Не успе+ла] [за Андре+ем] [затвориться две+рь], # [как я уви+дел] [в свое+м] [кабине+те]#
Не успе+ла[2] за Андре+ем затвориться **две+рь**[3], # как я уви+дел в свое+м **кабине+те**[3]#


[Высо+кого] [широкоплечего мужчи+ну], # [держа+вшего в одной руке] [бума+жный] [све+рток], #
Высо+кого[3] широкоплечего **мужчи+ну**[6], # **держа+вшего**[6] в одной руке бума+жный **све+рток**[6], #


[а в друго+й] – # [фура+жку] # [с кока+рдой]#
а в **друго+й**[3] – # **фура+жку**[6] # **с кока+рдой**[1]#

Рис. 4. Последовательность УМП предложения 4 (диктор А. Балакирев)


[Оде+т он] [со вку+сом] # [и по после+дней] [мо+де] # [в но+венький], #
Оде+т[6] он со **вку+сом**[1] # и по **после+дней**[6] **мо+де**[1] # **в но+венький**[6], #

[недавно сши+тый] # [трико+вый] [костю+м].#
недавно **сши+тый**[6] # трико+вый **костю+м**[1].#

Рис. 5. Последовательность УМП предложения 5 (диктор А. Балакирев)


[Лицо+] [ро+зовое], # [ру+ки] [велики+], # [гру+дь] [широ+кая], #
Лицо+[3] **ро+зовое**[1], # **ру+ки**[3] **велики+**[6], # **гру+дь**[3] **широ+кая**[6], #


[мускули+стая], # [во+лосы] [густы+], # [как у здоро+вого] [ма+льчика].#
мускули+стая[6], # **во+лосы**[3] **густы+**[6], # как у здоро+вого **ма+льчика**[2].#

Рис. 6. Последовательность УМП предложения 6 (диктор А. Балакирев)

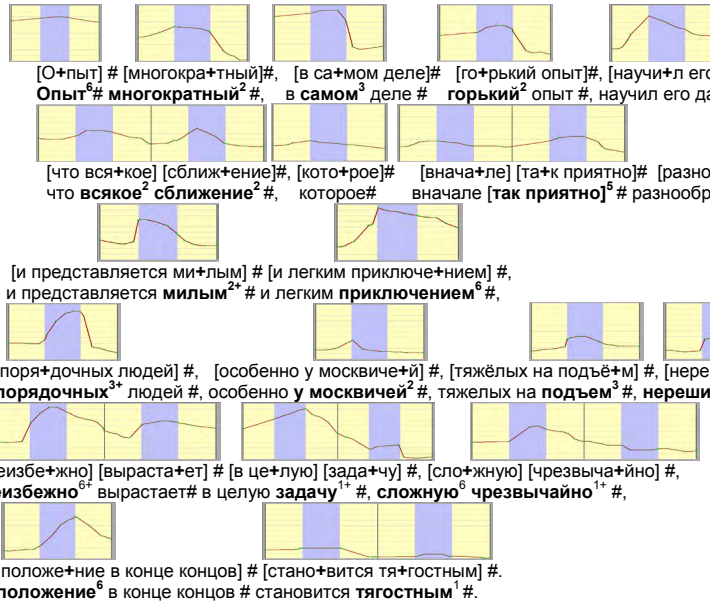


Рис. 7. Последовательность УМП предложения 7 (диктор С. Концевич)

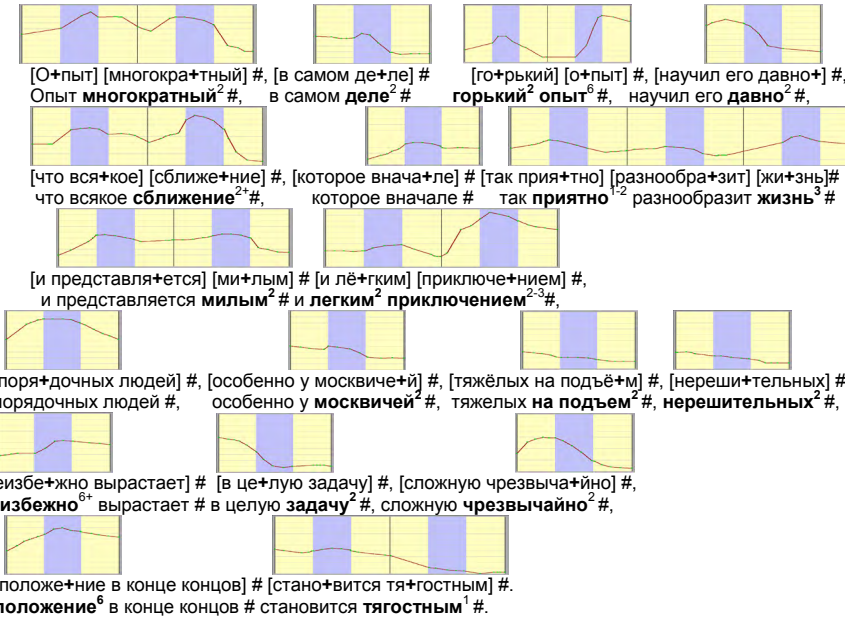


Рис. 8. Последовательность УМП предложения 7 (диктор В. Ветров)

3. Обсуждение результатов

В таблице 2 приведены сводные данные о просодической структуре семи исследованных повествовательных предложений в исполнении дикторов Александра Балакирева (А. Б.), Станислава Концевича (С. К.) и Владислава Ветрова (В. В.).

Таблица 2. Сводные данные о структуре исследованных предложений

№ предложения	Кол. слов	Кол. синтагм	Кол. АЕ	Кол. ИК (всего)	Кол. ИК1	Кол. ИК2	Кол. ИК3	Кол. ИК6
1 (диктор А. Б.)	32	9	19	9	1	1	2	5
2 (диктор А. Б.)	23	7	13	6	2	1	1	2
3 (диктор А. Б.)	20	9	13	7	1	1	4	1
4 (диктор А. Б.)	28	7	14	10	1	1	4	4
5 (диктор А. Б.)	15	5	8	7	3	0	0	4
6 (диктор А. Б.)	14	6	11	10	1	1	4	4
7 (диктор С. К.)	49	20	30	19	3	6	5	5
7 (диктор В. В.)	49	18	27	19	1	14	1	3
Всего:	230	81	135	87	13	25	21	28

Исходя из данных, представленных в таблице, можно сделать следующие общие количественные заключения о просодической структуре всей совокупности исследованных сложных повествовательных предложений:

1. Среднее количество слов в предложениях — 28, синтагм в предложениях — 10, слов в синтагмах — 2,8, АЕ в синтагмах — 1,6.
2. Отождествление АЕ как одной из возможных ИК, реализованной в речи диктора, осуществляется аудитором в 64%. Оставшиеся 36% не идентифицированы в рамках полного набора ИК1-ИК7.
3. Из полного набора 7-ми интонационных конструкций Е. А. Брызгуновой в рассмотренных повествовательных предложениях аудитором идентифицированы только 4 типа: ИК1, ИК2, ИК3, ИК6
4. Различными дикторами предпочтительно используются различные количественные наборы ИК1, ИК2, ИК3, ИК6. Более наглядно это проиллюстрировано на рис. 9, который показывает, что диктор Александр Балакирев преимущественно использует ИК6 и ИК3, диктор Станислав Концевич почти в одинаковой степени — ИК2, ИК3, ИК6, диктор Владислав Ветров в большинстве случаев использует ИК2.

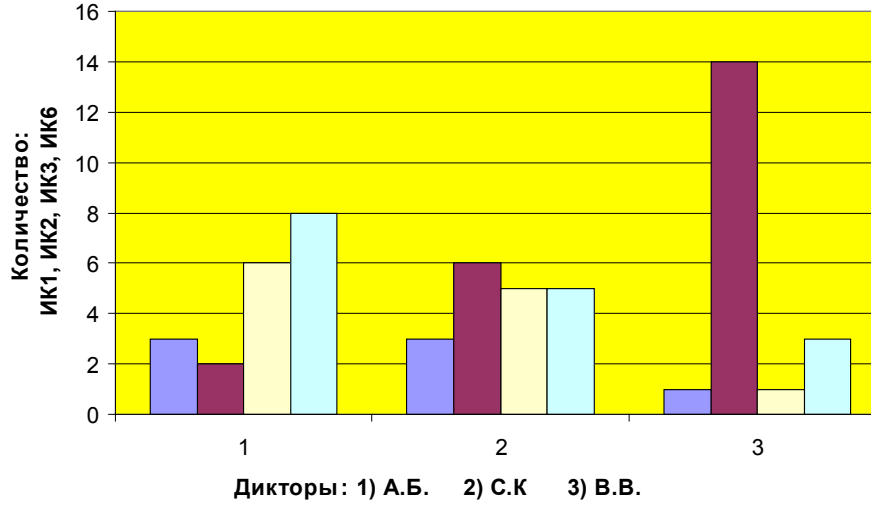


Рис. 9. Распределение количеств ИК1, ИК2, ИК3, ИК6, используемых 3-мя дикторами в процессе интонирования сложных предложений

Полученные в данной работе результаты базируются на предложенной ранее автором ПАЕ-модели (модели портретов акцентных единиц), эффективно используемой в интонационном блоке системы синтеза речи по тексту. ПАЕ-модель предполагает, что для определенного типа интонации топологические свойства мелодического контура АЕ не зависят от количественного и качественного содержания предъядра, ядра и заядра, и таким образом, она может быть представлена в нормированных координатах «Частота — Время». Это обеспечивает возможность представления семи интонационных конструкций Брызгуновой — {ИКi} — в виде набора их Универсальных Мелодических Портретов (УМП).

В работе [Лобанов, 2014] представлены «эталонные формы» УМП для различных ИК простых предложений, построенных на основе аудиоматериала из методического пособия по обучению РКИ [*Одинцова*, 2011]. Эти эталонные формы УМП для ИК1, ИК2, ИК3, ИК6 приведены на рис. 10 в верхнем ряду. В нижних рядах приведены УМП, экспериментально полученные для каждого из 3-х дикторов. Жирные кривые отображают средние значения УМП для множества их реализаций в анализируемых предложениях.

Из анализа формы кривых, представленных на рис. 10, можно сделать следующие выводы:

- Распределения кривых УМП ИК простых предложений (см. верхний ряд на рис. 10) значительно более компактны в сравнении с проанализированными сложными предложениями (см. нижние ряды);
- Форма средних значений кривых УМП ИК сложных предложений отличается от эталонных УМП ИК простых предложений в сторону меньшей выраженности их характерных свойств, хотя общие тенденции в поведении каждой из кривых УМП {ИК1, ИК2, ИК3, ИК6} в значительной степени сохраняются.

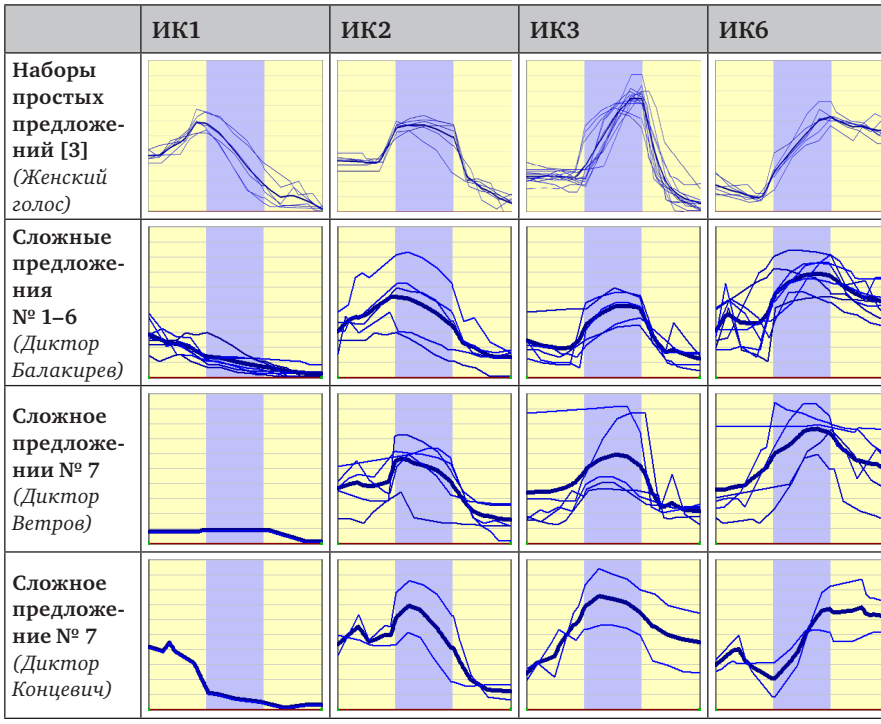


Рис. 10. Наборы экспериментальных УМП (ИК1, ИК2, ИК3, ИК6) устного чтения предложений различными дикторами

Заключение

В целом, исходя из результатов проведенного эксперимента, можно с определённой степенью уверенности утверждать, что в процессе устного чтения сложных предложений различными дикторами имеется тенденция к реализации определённых ИК, близких к эталонным ИК Брызгуновой, хотя их «выразительность» существенно ниже, чем при интонировании простых предложений. При этом в рассмотренных повествовательных предложениях из полного набора 7-ми типов ИК аудитором идентифицированы УМП только 4-х типов: ИК1, ИК2, ИК3, ИК6. С другой стороны, 36 % УМП оказались не идентифицированными в рамках набора из 7-ми интонационных конструкций Е. А. Брызгуновой. По-видимому, набор ИК для сложных предложений может оказаться более широким, чем классический набор ИК и его следует расширить. В заключение отметим, что в данной работе не ставилась задача анализа закономерностей использования той или иной ИК для маркировки последовательности АЕ текста. Решение такого рода вопросов представляется задачей дальнейших исследований. Автор надеется, что начальным побуждающим стимулом для

этого может послужить анализ последовательностей УМП различных предложений, приведенных на рис. 1–8.

Автор выражает глубокую благодарность Елене Карневской за методическую помощь в работе над первым этапом аудирования — членением на просодические синтагмы и маркировки АЕ, а также, в не меньшей степени, благодарность Татьяне Янко за проведение второго этапа аудирования — маркировку ИК.

References

1. *Bryzgunova E. A.* (1968) Sounds and Intonation of Russian Speech [Zvuki I Intonatsiya Russkoy Rechi], Science [Nauka], Moscow.
2. *Lobanov B. M., Okrut T. I.* (2014) Universal Melodic Portraits of Intonation Patterns in Russian Speech [Universal'nye Melodicheskie Portrety Intonacionnyh Konstruktsiy Russkoy Rechi], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2014" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2014"], Bekasovo, pp. 330–339.
3. *Odintsova I. V.* (2011) Sounds. Rhythmic. Intonation [Zvuki. Ritmika. Intonatsiya], Flinta-Science [Flinta-Nauka], Moscow.

АВТОМАТИЧЕСКАЯ КЛАССИФИКАЦИЯ РУССКИХ ГЛАГОЛОВ С ИСПОЛЬЗОВАНИЕМ ИНФОРМАЦИИ О МОРФО-СИНТАКСИЧЕСКОМ ОФОРМЛЕНИИ И СЕМАНТИЧЕСКИХ РОЛЯХ УЧАСТНИКОВ ФРЕЙМОВ

Ляшевская О. Н. (olesar@yandex.ru)[1,2],
Кашкин Е. В. (egorkashkin@rambler.ru)[2]

[1]Национальный исследовательский университет
Высшая школа экономики, Москва

[2]Институт русского языка им. В. В. Виноградова РАН, Москва

В статье описываются эксперименты по классификации русских глаголов на основе статистических данных, представленных в системе FrameBank (framebank.ru). Хотя лексикологи в основном отказались от мысли, что группы глаголов должны объединяться на основе способности к синтаксическим трансформациям (Apresjan 1967, Levin 1993), оценка близости контекстов по схожей дистрибуции лексики и синтаксических связей по-прежнему остается ведущим критерием для определения лексических типов. Компьютерная лингвистика заимствовала последний подход для получения глагольных классов для английского, немецкого и многих других языков (Dorr and Jones 1996; Lapata 1999; Schulte im Walde; Lenci 2014 и др.), строя векторы лексических и синтаксических признаков на основе корпусов текстов.

Наши эксперименты по семантической классификации русских глаголов базируются на статистике двух типов тегов, используемых в аннотации системы ФреймБанк, теге семантической роли и теге морфосинтаксического оформления участника. Поле глаголов речи было структурировано с помощью нескольких вариантов автоматической кластеризации на векторах; затем автоматические результаты мы сравнили с классификацией глаголов в словаре Л. Г. Бабенко (2007) и некоторыми другими построенными вручную классификациями. Классификация глаголов смены посессора была построена с помощью правил и затем была верифицирована относительно сети глагольных фреймов в англоязычной системе FrameNet. Проводится лингвистический анализ классификаций, получающихся только на морфосинтаксических признаках, только на признаках семантических ролей и классификаций на объединении этих признаков.

Ключевые слова: лексические классификации, глагол, FrameNet, FrameBank, семантические роли, морфосинтаксис, фреймовая семантика, лексикология, русский язык

INDUCING VERB CLASSES FROM FRAMES IN RUSSIAN: MORPHO-SYNTAX AND SEMANTIC ROLES[1]

Olga Lyashevskaya (olesar@yandex.ru)[1,2],
Egor Kashkin (egorkashkin@rambler.ru)[2]

[1]National Research University Higher School of Economics
[2]V. V. Vinogradov Russian Language Institute of RAS, Moscow

The paper presents clustering experiments on Russian verbs based on the statistical data drawn from the Russian FrameBank (framebank.ru). While lexicology has essentially abandoned the idea of syntactic transformations as the primary basis for grouping verbs into semantic classes (Apresjan 1967, Levin 1993), the hypothesis of the same lexical and syntactic distributional profiles underlying lexical clusters is still attractive. In computational linguistics, some attempts have been made to obtain verb classes for English, German and other languages using observable morpho-syntactic and lexical properties of context (Dorr and Jones 1996; Lapata 1999; Schulte im Walde 2006; Lenci 2014, among others).

Our experiments on semantic classification of Russian verbs are based on two types of tags embedded in the annotation of argument constructions: a) semantic roles and b) morpho-syntactic patterns. The domain of speech verbs is classified automatically on vectors, and the resulting clusters are contrasted against Babenko (2007)'s semantic classes and three other manual classifications. The classes within the domain of possessive verbs are constructed using rule-based solutions and evaluated against Berkeley FrameNet verb clusters. We conclude that clustering on morpho-syntactic (pure formal) patterns loses the race to more intelligent approaches which take into account semantic roles.

Key words: lexical classifications, verb, FrameNet, FrameBank, semantic roles, morpho-syntax, frame semantics, lexicology, Russian language

1. Introduction

The systematic lexicography approach (Apresjan 2000, 2002) generalizes over words according to their properties to share patterns of conceptualization and the regular paths of meaning development and interaction (polysemy, antonymy, etc). Word classes are also expected to manifest similar trends in grammatical, syntactic and lexical co-occurrence behavior. These ideas establish grounds for the unified

[1] This work was partly supported by the Russian Foundation for the Humanities, project #13-04-12020 "New Open Electronic Thesaurus for Russian" and Russian Basic Research Foundation, project # 15-07-09306 "Evaluation benchmark for information retrieval".

representation of word classes or 'types' (in terms of their semantics and 'lexical grammar') in lexicography, functional and cognitive linguistics, and computational linguistics.

In computational linguistics, word classes and nets help to reduce the sparseness of lexical vectors which measure how often each word from the corpus (one coordinate axis) occurs in the context of the target word. If two words belong to the same class, the corresponding dimensions can be collapsed into one; this can also help to associate context elements not available in the training set. As a result, the use of lexical classes (along with other categories available in corpus annotation such as parts of speech, lemmas, ie. the sets of word forms, semantic roles, syntactic relation types etc.) affects data pruning, feature weighting and feature selection and can be considered potentially good way to improve machine learning. The only limitation is the availability of large-scale lexical classifications as open-access resources.

There is a number of manual builds of verb classes for several languages, including English (Levin 1993 and its implementation in VerbNet, Palmer 2009; Baker, Fillmore, and Lowe 1998), Spanish (Vázquez et al. 2000), Russian (Babenko 2007, Shvedova 1998–2007), etc. More attempts have been made to obtain verb classes automatically (Dorr and Jones 1996; Lapata 1999; Korhonen 2002; Schulte im Walde 2006; Lenci 2014, etc). Lenci (2014) distinguishes between the ontology-based and distribution-based classifications. As an instance, the verbs *eat* and *devour* belong to the same group in the ontolology-based classification since they evoke the same frame Ingestion 'an Ingestor consumes Ingestibles'; in contrast, *eat* and *devour* do not share certain syntactic properties such as object drop and conative construction and therefore can be placed in different groups in at least some versions of distribution-based classifications. Lenci's example is misleading since there are two context vector models underlying the distribution-based classification. If the idea of syntactic transformations is taken into account, then the target words are seen as being in two states (cf. two isotopes of a chemical element) in which they behave differently and their context image consists basically of two classes of vectors. In the more straightforward reading of the distributional hypothesis, the context vectors of the target word form a homogeneous image. The transformational hypothesis has been put under question by Construction Grammar and quantitative corpus-based approaches. As corpus data show, the alternations are rather peripheral than central phenomenon (see discussion in Kuznetsova, Lyashevskaya 2009; Kuznetsova 2013), and verbs from the same lexical class demonstrate strong statistical preferences for either one or another alternating construction (Gries, Stefanowitch 2004). Therefore, we leave transformations out of the model in order to make it less computationally complex.

In our approach, we take both latent frame-based cues and observable morpho-syntactic cues in order to evaluate their classification strength in the task of Russian verb clustering. The paper is structured as follows. Section 2 outlines Russian FrameBank as a data source for our case studies. In Section 3, we introduce the case study on speech verbs clusters which were classified by machine learning and contrasted against four gold standards. Section 4 summarizes an experiment where possessive verbs were classified using rule-based solutions and then evaluated against Berkeley FrameNet verb clusters. Section 5 concludes.

2. Data

The Russian FrameBank (framebank.ru, Lyashevskaya, Kuznetsova 2009; Kashkin, Lyashevskaya 2013; Lyashevskaya, Kashkin 2014) includes a dictionary of lexical constructions and a corpus of manually annotated sentences (up to 100 examples from the Russian National Corpus for each target word). In our experiments we use two types of tags embedded in the annotation of argument constructions both in the dictionary and the corpus: a) semantic roles and b) morpho-syntactic properties which form a formal pattern of constructions. For example, the verb *govorit'* is associated with a number of frames, cf. the frame of CONVERSATION:

Semantic roles of frame elements: Speaker, Counter-agent, Topic

Furthermore, each frame is associated with a set of lexical constructions, cf. The two-argument construction in (1) with a particilar pairing of meaning (formalized as a combination of semantic roles) and form (formalized as a set of morpho-syntactic constraints):

(1) *Dmitriev govoril s dochkoy*. 'Dmitriev talked to his daugher'.
Semantic role pattern: <Speaker, Counter-agent>
Morpho-syntactic pattern: <NPnom[2], *s* 'with' + NPgen>

Example (2) presents the frame of INFORMATION TRANSFER (e. g. saying smth. to smb.) and the three-argument construction of the verb *govorit'* 'say':

(2) — *Vsego etogo nedostatochno, — govoril mne Dviniatin*. 'Dviniatin said to me, 'All this is not enough'.
Semantic role pattern: <Speaker, Addressee, Message-as-content>
Morpho-syntactic pattern: <NPnom, NPdat, CL>

In our first case study we explored the contexts of speech verbs which were assigned to frames where at least one participant plays a role of Speaker, Addressee, Topic, or Message-as-content. The data set included vectors for 80 speech verbs having speech frames being associated with their primary meaning.

For the second case study we used only data from the dictionary database, namely, information on semantic roles, morpho-syntactic tags, and their matches in lexical constructions. We included into our experiment 128 verbs having the arguments with the roles of Initial Possessor or Eventual Possessor. If a verb represented such role patterns for more than one frame in the database (e. g., the verb *vz'at'* which may refer either to TAKING or to BYING), these cases were counted as different verbs (e. g., we analyzed a verb *vz'at' 1* 'to take' and the verb *vz'at' 9* 'to buy').

---

[2] Here and throughout, *nom* stands for Nominative case, *gen* for Genitive, *dat* for Dative, *acc* for Accusative, *ins* for Instrumental, *loc* for Locative, *CL* for clause; {ADV / PRfrom_where + Npx} refers to any prepositional phrase or adverb with the meaning 'from a certain source'.

3. Case study 1: speech verbs

In the first case study, we explore subclasses of speech verbs. These verbs differ in terms of associated morpho-syntactic constructions and combinations of participants in the frames they evoke; the most common set of roles usually includes 1) Speaker; 2) Addresse or Counter-Agent; 3) Message-as-Topic and / or Message-as-content. Rare cases include such roles as Motivation (cf. *khvalit' za pirogi* 'to praise (smb.) for cakes'), Quantity (cf. *povtorit' dvazhdy* 'to repeat twice'), Point of Destination (cf. *zvat' v park* 'to call (smb.) to the park'), etc.

Figures 1 and 2 report how often morpho-syntactic and role tags occur in the context of verbs (in each case the verbs are sorted separately according to the ratio of contexts that include a given element).

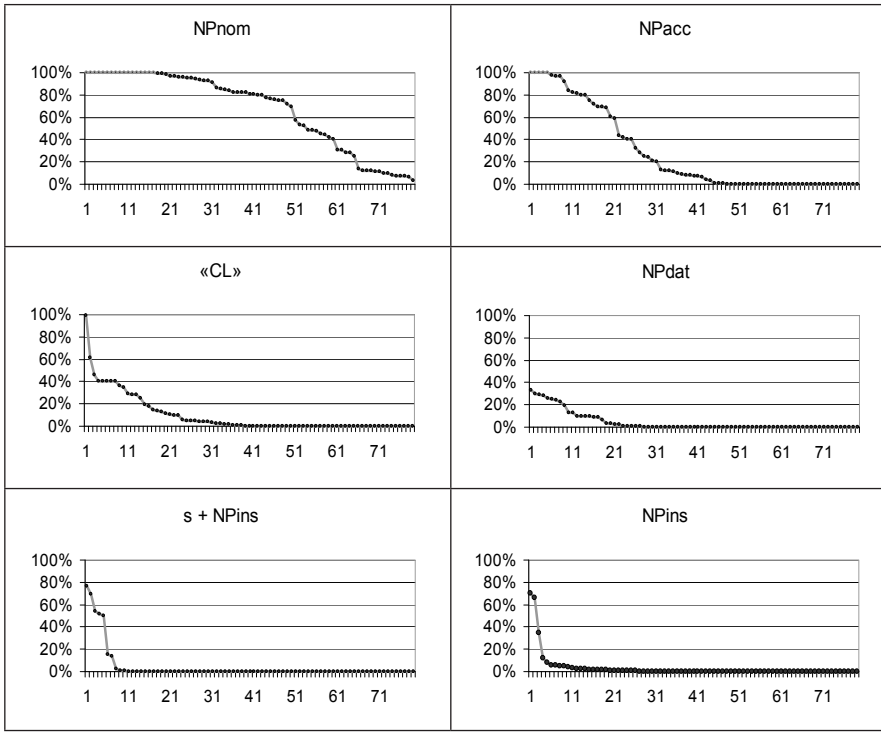


Fig. 1. Verbs sorted by the ratio of top-6 morphological tags[3] in their context, in % of tagged examples for each verb

[3] *s* + S ins is a prepositional group which means "with + S ins', «CL» stands for the direct speech clause; see also footnote 2.

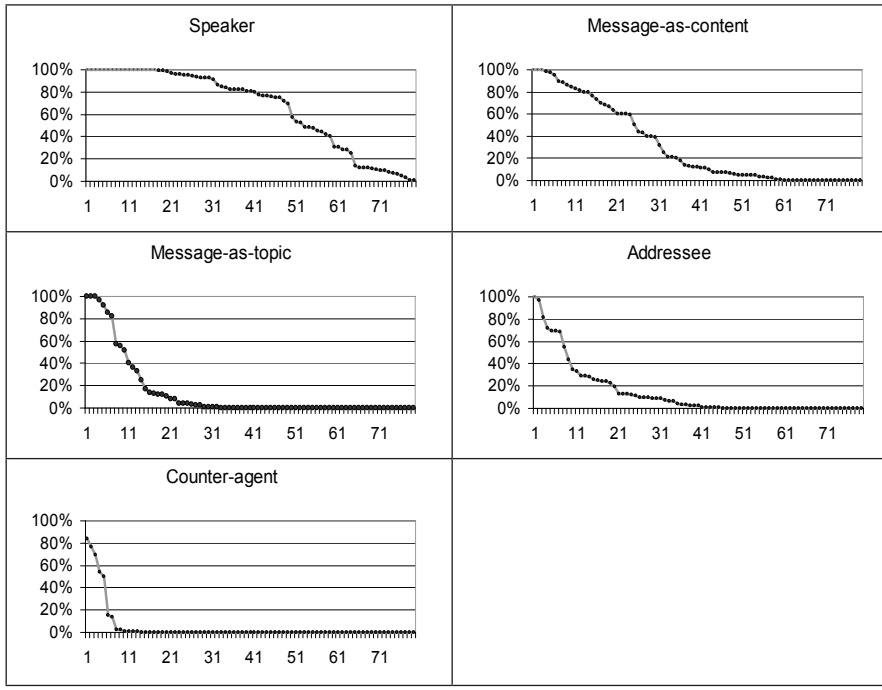


Fig. 2. Verbs sorted by the ratio of top-5 semantic role tags in their context (Speaker, Message-as-content, Addressee, Message-as-topic, Counter-agent), in % of tagged examples

Three types of vectors presenting morpho-syntactic tags (e. g. *o* + NPloc), semantic roles (e. g. Topic), or their matches (e. g. *o* + NPloc|Topic) are gathered taking frequencies from annotated corpus as coordinates. The data set includes vectors for 80 verbs which refer to speech in their primary meaning. As a result, there is a 34-dimensional vector space for morhosyntactic tags (tags that occur less than 5 times such as *v kachestve* + NPgen 'qua', *ot imeni* + NPgen 'on behalf of' are removed from the data set), 20-dimensional space for semantic roles (roles that occur less than 5 times such as Result and Direction are removed as well), and a 71-dimensional space for the combined features of morpho-syntax and semantic roles (also pruned with the threshold of 5).

Table 1 shows the comparison of k-means-based clustering[4] results against four variants of gold standard. The metrics of Purity (PU), Collocation (CO), and F1 are understood in accordance with (Lang, Lapata 2011). PU is calculated as (3), where

[4] K-means is a traditional algorithm which finds the best partition of points in n-dimensional vector space (in our case, verbs) into k clusters such that the squared error between the empirical mean of a cluster and the points in the cluster is minimized (for an overview and discussion, see Jain 2010). K is a fixed positive integer number specified by the researcher. K-means starts with a (random) initial partitition with K seed points selected as cluster centers and initial assignment of data points to clusters. After that, the data points are reassigned to its closest cluster center and then new cluster centers are calculated, and these two steps are repeated until cluster membership stabilizes.

n denotes the total number of instances, G_j is the set of instances belonging to the *j*-th gold class and Ci is the set of instances belonging to the i-th cluster. CO measures how well the procedure meets the goal of clustering all gold instances with the same label into a single predicted cluster and is computed according to (4). F1 is the harmonic mean of Pu and CO.

$$\mathrm{Pu} = \frac{1}{n}\sum_{i=1}^{n_C} \max_{j=1,\ldots,n_G} \left|C_i \cap G_j\right| \tag{3}$$

$$\mathrm{Co} = \frac{1}{n}\sum_{j=1}^{n_G} \max_{i=1,\ldots,n_C} \left|C_i \cap G_j\right| \tag{4}$$

It is important to keep in mind that there cannot be an ideal gold classification due to different principles that could be applied to data (e. g. thematic proximity, event structure, pragmatic goals, etc., cf. the classifications of Wierzbicka 1983; Bogdanov 1990; Glovinskaya 1993; Era Kuznetsova 1989; Shvedova 1998–2007; Babenko 2007, among many others). Rather, it is better think of probability to which a linguistic community would agree to assign the verb C to the same cluster as A and B. Given that, we built four variants of classification: 1) based on Babenko 2007's classes (our verbs fall into 23 Babenko's classes including six classes of speech verbs and some classes outside the speech domain like behavior, emotion, etc.); 2) based on the role of the 2nd participant: addressee-like verbs, counteragent-like verbs, patient-like verbs, benefactor-like verbs, no-addressee verbs; 3) based on the goals of the speaker (7 classes); 4) its more detailed version with 19 classes of sharing information, getting information, symmetric communication, and various types of speech affect like asking, abusing, etc. The number of verbs accumulated in classes in each gold standard is reported in Table 2.

According to F-measure, Roles generally overperform Forms with the only exception of the best split in cross-validation against the last gold standard (Goal33). Interestingly, Forms perform better at smaller *k*-s while Roles work better at larger *k* values. Roles & Forms optimizes the split in three cases of four but there are many cases there Roles demonstrate higher scores than Roles & Forms at the same *k* value. Thus, we conclude that the hierarchy of features predicting speech verb classes looks like the following: Roles & Forms ≥ Roles > Forms.

In the second trial, we use the same vector datasets (Forms, Roles, Roles & Forms) to compare three hierarchical cluster trees based on cosine distances and to follow the verbs changing (or not changing) their position in clusters. At $k = 7$, there are eight clusters of size 4 to 14 where the verbs group together under all three conditions (e. g. {*besedovat', zdorovat'sja, obschat'sja, prostit'sja*}, {*blagodarit', informirovat', pozdravit', privetstvovat', khvalit', zvat', klikat', oprosit'*}, etc.; 63 verbs in total). Due to the small number of verbs in clusters, it is easy to inspect the homogeneity of clusters manually. The indisputable errors include such pairings as *obvinit'* and *ugovorit'* ('accuse' and 'persuade'), *prokl'ast'* and *obosnovat'* ('imprecate' and 'justify'), *podkhvatit'* and *ugrozhat'* ('play along' and 'threaten').

Table 1. Evaluation of k-means clustering by Purity (PU), Collocation (CO) and their harmonic mean F1 (Lang, Lapata 2011). The best trials are bold-faced. PU tends to increase and CO tends to decrease as the number of clusters *k* increases

	Babenko (k(gold)=23, elements in the largest class = 26)								
	Roles & Forms			**Roles**			**Forms**		
k	PU	CO	F1	PU	CO	F1	PU	CO	F1
4	0.3418	**0.9114**	0.4971	0.3544	**0.8861**	0.5063	0.3418	**0.9241**	**0.4990**
5	0.3544	**0.9114**	0.5104	0.3797	0.6709	0.4850	0.3418	0.7342	0.4664
13	0.4430	0.6329	**0.5212**	0.4557	0.5949	0.5161	0.4304	0.5570	0.4856
20	0.4810	0.4937	0.4873	0.5316	0.5190	0.5252	**0.4557**	0.4430	0.4493
21	**0.5063**	0.5190	0.5126	0.5316	0.5190	0.5252	0.4430	0.4810	0.4612
23	0.4810	0.4684	0.4746	0.5570	0.5063	**0.5304**	**0.4557**	0.4430	0.4493
25	0.4304	0.4684	0.4486	**0.5823**	0.4557	0.5113	0.4177	0.4304	0.4240
	Role2 (k(gold)=7, elements in the largest class = 33)								
	Roles & Forms			**Roles**			**Forms**		
k	PU	CO	F1	PU	CO	F1	PU	CO	F1
4	0.4051	**0.9114**	0.5609	0.4430	**0.8228**	**0.5759**	0.4051	**0.9241**	**0.5632**
5	0.4177	**0.9114**	0.5729	0.4937	0.5949	0.5396	0.4051	0.6076	0.4861
7	0.5570	0.6582	**0.6034**	0.4430	0.5696	0.4984	0.4937	0.5063	0.4999
10	0.4937	0.6076	0.5447	0.5696	0.5190	0.5431	0.4684	0.4304	0.4486
	SpeakerGoals ((k(gold)=7, elements in the largest class = 30)								
	Roles & Forms			**Roles**			**Forms**		
k	PU	CO	F1	PU	CO	F1	PU	CO	F1
4	0.4304	**0.9114**	0.5847	0.4557	**0.8354**	0.5897	0.4304	**0.9241**	**0.5872**
5	0.4430	**0.9114**	0.5962	0.5190	0.5696	0.5431	0.4430	0.6076	0.5124
7	**0.5823**	0.6962	**0.6342**	0.4810	0.5316	0.5051	0.4937	0.5190	0.5060
8	0.5696	0.5570	0.5632	**0.6456**	0.5443	**0.5906**	0.4810	0.5063	0.4933
9	0.5696	0.5570	0.5632	0.6203	0.4810	0.5418	**0.5316**	0.4177	0.4678
	Goals19 (k(gold)=19, elements in the largest class = 26)								
	Roles & Forms			**Roles**			**Forms**		
k	PU	CO	F1	PU	CO	F1	PU	CO	F1
4	0.3418	**0.9241**	0.4990	0.3671	**0.8354**	0.5101	0.3418	**0.9241**	0.4990
5	0.3544	**0.9241**	0.5123	0.3924	0.6582	0.4917	0.3418	0.6835	0.4557
11	0.3924	0.6076	0.4768	0.4684	0.5823	0.5191	0.4810	0.6709	**0.5603**
19	0.5316	0.6076	**0.5671**	0.5443	0.5190	0.5313	0.5190	0.5063	0.5126
22	**0.5696**	0.4684	0.5140	0.5823	0.4937	0.5343	0.5316	0.4557	0.4907
23	0.5443	0.4430	0.4885	**0.5949**	0.4810	0.5319	0.5316	0.4430	0.4833
24	0.5443	0.4810	0.5107	**0.5949**	0.4937	**0.5396**	**0.5443**	0.4430	0.4885
25	**0.5696**	0.4557	0.5063	**0.5949**	0.4684	0.5241	**0.5443**	0.4430	0.4885

Table 2. The size of classes in four gold standards, sorted from larger to smaller groups

Babenko	26	10	9	7	4	4	2	2	1	1	1	1 in each other class
Role2	30	20	16	9	2	1	1					
SpeakerGoals	33	25	8	5	3	3	2					
Goals19	26	9	8	6	5	4	3	3	3	2	2	1 in each other class

For example, *obvinit'* and *ugovorit'* are similar in terms of Roles vectors (Speaker—Addressee—Message-as-content available in context), but partly different in terms of their main morpho-syntactic patterns, cf. NPnom NPacc *v*+NPloc and NPnom NPacc VPinf. The cosine measure shows very small distance between their vectors due to orthogonality effects where the following situation takes place:

dimensions	1	2	3	4	5	6	7	8	*i*
vector 1	a	b	-	c	-	-	-	1	-
vector 2	a	b	c	-	-	1	-	-	-

In this particular case, the number of *v*+NPloc and VPinf tags which occur in context is 55 and 54, respectively, while other tags (e. g. Message-as-content|Conj + CL, reason|Sins, etc.) occur not more than 1 time. If the verb is somewhat unique (cf. the high ratio of *v*+NPloc tags in the context of *objasnit'*), the probability that the hierarchical clustering will produce an error becomes even greater.

4. Case study 2: possessive verbs

Our second case study deals with verbs that refer to change of possession (frames of buying, stealing, giving, etc). We created a boolean table with the verbs in the rows and with the possible pairings of morpho-syntactic tags and semantic roles (e. g. NPnom|Eventual Possessor) in the columns. It was marked in the table which clusters of morpho-syntactic patterns and semantic roles are compatible with each verb.

Further, all the verbs were manually divided into clusters based on a set of heuristics involving their morpho-syntactic patterns and semantic roles. The results were verified with an agglomerative clustering method[5]. As a result, we found four main clusters of verbs, most of them are further subdivided into several smaller clusters. The structure of the possessive domain is represented below (the verbs of each subclass are also included into all the parent classes). The figures in brackets show the number of verbs which have fallen into a particular class.

1. Take (34 verbs): verbs with patterns where NPnom is an Eventual Possessor (e. g. *brat' 1* 'to take', *otn'at' 1* 'to take away').

1.1 Buy (4 verbs): verbs having a pattern NPnom V NPacc *za* + NPacc (Eventual Possessor—Patient—Price)[6], e. g. *kupit'* 1 'to buy', *ar'endovat' 1* 'to rent'

1.2 Steal (10 verbs): verbs having a pattern NPnom V NPacc {ADV / PRfrom_where + NPx} (Eventual Possessor—Patient—Starting Point). Interestingly, they

[5] hclust() method in R, package 'stats', see Langfelder, Horvath 2012. The agglomerative starts with one cluster for each verb and merges the pair of clusters with the minimum intercluster distance.

[6] The verbs of Buying may obviously occur in some other patterns, e. g. in those expressing an Initial Possessor or a Place. However, these patterns are not specific for this class of verbs and cannot serve as a diagnostics for it, so we do not refer to them when defining the class of Buying. The descriptions of the other classes follow the same principle: what is specially mentioned is only the diagnostic patterns for each class.

tend to describe the events of theft[7]. There are 10 verbs with this pattern in the database, and 7 of them obviously refer to stealing: *vorovat' 1* 'to steal', *krast' 1* 'to steal', *pohitit'* 1 'to steal, to kidnap', *taskat' 4* 'to pinch', *taščit'* 6 'to pinch', *t'anut' 11* 'to swipe', *uv'esti 2* 'to steal (usually cattle or a car)' The other three verbs with this formal pattern are *brat' 1* 'to take', *zabrat' 1* 'to take away', and *hvatat' 1* 'to snatch', which may all be used in neutral possessive contexts outside the domain of stealing. Thus, our protocol suggests here a broader result than necessary, but it is important that all the verbs of stealing are inside this formal class, the only possible "lost" verb is *otn'at' 1* 'to take away, to deprive'

1.3 Receive (3 verbs): verbs having a pattern NPnom V NPacc *ot* + NPgen (Eventual Possessor—Patient—Initial Possessor). There are three verbs—*polučit' 1* 'to receive', *prin'at' 1* 'to accept (e. g. a present)', *prin'at' 4* 'to accept (e. g., an advertisement)', — where the Initial possessor is marked not by the more frequent PP *u* + NPgen, but by *ot* + NPgen. This opposition seems to be related to the degree of agentivity: verbs with *u* + NPgen like *brat' 1* 'to take' imply a more active behavior of the Agent than verbs with *ot* + NPgen argument.

1.4 Earn (2 verbs): verbs expressing Price as NPacc. This is a subclass of taking verbs which corresponds to the events of earning money. In FrameBank these are the verbs *zarabotat' 1* and *polučit' 2* meaning 'to earn (money)'.

1.5 Borrow (3 verbs): verbs of taking which have Time Period among their participants. They refer to the events of borrowing: *ar'endovat' 1* 'to lease sth. from sb.', *zan'at' 8* 'to borrow', *sn'at' 13* 'to rent'

2. Give (90 verbs): verbs with patterns where NPnom is an Initial Possessor, cf. *dat' 1* 'to give', *vozvratit' 1* 'to return sth. to sb.', *pr'epodn'esti 1* 'to present sth. to sb.', etc.

2.1 Sell (6 verbs): verbs having a pattern NPnom V NPacc *za* + NPacc (Initial Possessor—Patient—Price), e. g. *prodat' 1* 'to sell', *sdat' 4* 'to lease sth. to sb.'

2.2 Pay (2 verbs): verbs of giving (*platit' 1* 'to pay', *ustupit' 3* 'to take off (a price)') which have a direct object expressing Price, similarly to the verbs of earning.

2.3 Give somewhere (8 verbs): verbs having a pattern NPnom V NPacc {ADV / PRwhere(to) + NPx} (Initial Possessor—Patient—Point of destination). We put them together under a technical label "Give somewhere" This subclass doesn't appear to be homogenous, but the verbs included there follow some semantic tendencies. First, these are the verbs *vernut' 1* and *vozvratit' 1* meaning 'to return sth. to sb.' Second, this subclass includes the verbs *vyslat' 1*, *poslat' 2* both meaning 'to send' and *p'er'edat' 1* 'to pass sth. to sb.' which presume an intermediary in the change of possession. Third, there are verbs *podat' 2* 'to submit', *sdat' 1* 'to return, to surrender', and *sdat' 2* 'to submit, to hand in' also belonging to this subclass and referring to change of possession as a part of social relationship between the Initial Possessor and some kind of authorities being the Eventual Possessor.

---

[7] Here our results are in line with Apresjan 1967: 176–177, where the class of theft is singled out on the basis of its constructional properties.

2.4 Give with some goal (12 verbs): verbs having a pattern NPnom V NPacc *na* + NPacc (Initial Possessor—Patient—Goal) or a pattern NPnom V NPacc NPdat *na* + NPacc (Initial Possessor—Patient—Eventual Possessor—Goal). All of them conceptualize giving as an action intended to achieve some goal, cf. *assignovat'* 1 'to allocate', *žertvovat' 1* 'to donate', *tratit' 1, 2* 'to spend', *darit' 1* 'to make a present', *pr'epodn'esti 1* 'to present (a gift)' etc.

2.5 Supply (8 verbs): verbs (*balovat' 2* 'to make sb. glad by giving sth.', *vooružit' 1* 'to arm', *nagradit' 1* 'to award', *ob'esp'ečit' 1* 'to provide', *(n'e) obid'et' 2* 'not to stint sb. of sth.', *obogatit' 2* 'to enrich', *ssudit' 1* 'to loan') with a pattern NPnom V NPacc NPins (Initial Possessor—Eventual Possessor—Patient), a verb *od'et' 2* 'to provide clothes for sb' with a pattern NPnom V NPacc (Initial Possessor—Eventual Possessor), and a verb *obogatit' 1* 'to enrich' with a pattern NPnom V NPacc (Method—Eventual Possessor). The core of this set (*vooružit' 1*, *ob'esp'ečit' 1, obogatit' 1, 2, ssudit' 1, od'et' 2*) describes supplying sb. with sth. necessary. However, there are three peripheral verbs (*balovat' 2, nagradit' 1, (n'e) obid'et' 2*) falling into this subclass.

2.6 Lend (2 verbs): verbs of giving which have Time Period among their participants (*odolžit'* 1 'to lend', *sdat' 4* 'to rent out').

3. Exchange (1 verb): a verb *men'at's'a 1* 'to exchange', as its NPnom is Possessor

4. Other types

4.1 Owe (1 verb): a verb *sl'edovat' 9* 'to owe (lit.: to follow)' with its specific patterns, e. g. *Skol'ko* (ADV) *s nih* (s + NPgen) ***sl'edujet*** *za r'emont* (za + NPacc) 'How much do they owe for the repair (lit.: How much follows from them for the repair)?'

4.2 Go to somebody (2 verbs): two verbs with patterns where NPnom is a Patient—*dostat's'a 1* and *otojti 10*—meaning 'to go to sb.'

As has been stated above, our classification of the verbs is based both on their morpho-syntactic patterns and on the sets of semantic roles. If treated separately, these two criteria appear to be less fruitful than their combination. The semantic roles without the syntactic patterns fail to produce an adequate classification, since most possessive verbs are conversives (in the terminology of Apresjan 1974/1995: 256–283) and involve the same set of participants getting different syntactic ranks. The morpho-syntactic structure is more successful for clustering the possessive domain, e. g. the verbs of Taking and Giving complementary fit the patterns Snom V Sacc *u* + Sgen and Snom V Sacc Sdat respectively, the pattern Snom V Sacc *ot* + Sgen is unique for the verbs of Receiving. In many cases, however, the syntactic clustering lacks information on semantic roles and therefore produces too broad classes (as is sometimes the case in Apresjan 1967, along with a great deal of reliable correlations between semantic classes and constructional patterns). Thus, the classes of Buying and Selling admit the same syntactic pattern Snom V Sacc *za* + Sacc and are differentiated due to different correspondence between the semantic roles and syntactic participants. The classes of Paying and Earning encounter a similar problem, being marked out on the grounds

of Price being their direct object. In the pattern Snom V Sacc *na* + Sacc, the PP may perform 7 different roles in different verbs (Patient, Resource, Period, Eventual Possessor, Point of Destination, Price, Goal), so its specification as Goal is necessary for defining the class of "Giving with some goal". The patterns Snom V Sacc Sins and Snom V Sacc {ADV / PRwhere(to) + Sx} mostly correspond to the classes of Supplying and "Giving somewhere", but the syntactic clustering without the semantic roles produces 2 and 1 false results respectively (including *brat' 1* 'to take' and *kupit' 1* 'to buy' into the former domain and *vyslat' 1* 'to send' into the latter).

We have compared our results with the data of Berkeley FrameNet as the gold standard. The latter contains a frame of Giving with 6 subframes (Commerce_pay, Commerce_sell, Lending, Submitting_documents, Supply, Surrendering_possession), and the frame of Getting with 8 subframes (Amassing, Commerce_buy, Commerce_collect, Kidnapping, Receiving with the subframe of Borrowing, and Taking, further inherited by Theft).

The basic distinction between Giving and Getting is the same in FrameNet and in our survey. The frames of Commerce_pay, Commerce_sell, Lending, Commerce_buy, Receiving, Borrowing, Taking, and Theft transparently correspond to our classes. The frame of Amassing (e. g., *Bogs accumulate carbon for thousands of years*) is outside the possessive domain in FrameBank. The domain of kidnapping seems to be much more elaborated in English than in Russian, including even such specific verbs as *to shanghai* defined in the Oxford Dictionary as 'to force (someone) to join a ship lacking a full crew by drugging them or using other underhand means', therefore we haven't revealed this verb class in FrameBank. The frame of Commerce_collect has a bit strange definition in FrameNet ('Subframe of Commerce_money-transfer in which the Seller comes to have the Money', e. g. *The man at the counter collected payment from Lee for his dry-cleaning*), as the grounds for focusing on the motion of the Seller are not quite clear, but it roughly corresponds to our subtype "Earn"

The frames of Submitting_documents and Supply are both included into broader classes with some periphery ("Give somewhere" and "Supply", respectively). The only frame present in FrameNet but missing in our clustering is Surrendering_possession ('A Surrenderer is compelled to transfer a Theme to a Recipient', e. g. *Shortly after the boy surrendered the gun, the three remaining warriors made a rush for liberty*): we suggest that these verbs are treated separately due to rather fine-grained semantic reasons and do not seem to have their specific constructions.

Interestingly, our protocol has shown the subclass of Giving with some goal ('to allocate', 'to donate', etc.). These verbs do not form a single class in FrameNet, but intuitively they form a homogenous semantic class with a common set of participants, therefore our method seems to have been more successful here.

5. Conclusions

There is a great many statistical approaches to clustering word vectors which have been developed over the past decades. With access to ever growing corpora and handy script libraries and ready-made services, the task of clustering verbs

in different languages and domains seems pretty straightforward. As Alessandro Lenci pointed out in the draft of his paper (Lenci 2014), 'We have no doubt that verbs can be grouped into classes, since almost everything can be classified'. In this paper we have argued for the need to draw more attention to the input data structure while using the same algorithms and to involve the multiple gold standard approach.

Our case studies of speech verbs and possessive verbs have shown that unsupervised clustering performs better if semantic roles are taken into account, either as the only input (in the case of speech verbs) or together with the morpho-syntactic patterns. These observations have been made on rather small experimental dataset available for Russian and imply that the future development of this approach would require enhancing large corpora with SRL annotations. Given the recent success in deep learning and semantic role labeling in general (see Lang, Lapata 2014; Hermann et al. 2014; Täckström et al. 2015 etc.) and in Russian SRL parsing (Smirnov et al. 2014; Kuznetsov 2015), this does not sound as an unrealistic challenge.

References

1. *Apresjan, Juri.* 1967. Experimental'noe issledovanie semantiki russkogo glagola. Moscow.
2. *Apresjan, Juri.* 2000. Systematic lexicograpy. Oxford: Oxford University Press.
3. *Apresjan, Juri.* 2002. Principles of systematic lexicography. In Marie-Heléne Corrèard (ed.), Lexicography and Natural Language Processing. A Festschrift in Honour of B. T. S. Atkins. Euralex, Grenoble, pp. 91–104.
4. *Apresjan, Juri.* 1995. Integral'noe opisanie iazyka i sistemnaia leksikographiia [An Integrated Description of Language and Systematic Lexicography]. Moscow: Jazyki russkoj kul'tury.
5. *Apresjan, Jury D., Pall, Erna.* 1982. Russian verb—Hungarian verb. Government and combinability [Russkij glagol—vengerskij glagol. Upravlenie i sochetaemost'], Tankyonvkiado, Budapest.
6. *Babenko, Ludmila G.* 2007. Big Explanatory Dictionary of Russian Verbs: Ideographic description. Synonyms. Antonyms. English translation equivalents [Bol'shoj tolkovyj slovar' russkix glagolov: Ideograficheskoe opisanie. Sinonimy. Antonimy. Anglijskie Ekvivalenty]. Moscow: AST-press.
7. *Baker, Collin F., Charles J. Fillmore, and John B. Lowe.* 1998. The Berkeley FrameNet Project. In Proceedings of the 17th International Conference on Computational Linguistics and the 36th Annual Meeting of the Association for Computational Linguistics, Montreal, Canada. Pp. 86–90.
8. *Bogdanov V. V.* 1990. Speech communication: pragmatic and semantic aspects [Rechevoe obschenie: pragmaticheskie i semanticheskie aspekty]. Leningrad: LGU.
9. *Dorr, Bonnie J., Jones, Doug.* Role of word sense disambiguation in lexical acquisition: Predicting semantics from syntactic cues. Proceedings of the 16th conference on Computational linguistics-Volume 1. Association for Computational Linguistics, 1996.

10. *Glovinskaya, Marina Ya.* 1993. Semantics of speech verbs from the point of view of the speech act theory [Semantika glagolov rechi s tochki zrenija teorii rechevykh aktov]. In The Russian language in its functioning [Russkiy jazyk v ego funkcionirovanii]. Moscow. Pp. 158–218.
11. *Gries, Stefan Th. & Anatol Stefanowitsch.* 2004. Extending collostructional analysis: a corpus-based perspective on 'alternations'. International Journal of Corpus Linguistics 9 (1). Pp. 97–129.
12. *Hanks, Patrick.* 1996. Contextual Dependency and Lexical Sets. International Journal of Corpus Linguistics, Vol. 1(1), pp. 75–98.
13. *Hermann, Karl Moritz, Dipanjan Das, Jason Weston, and Kuzman Ganchev.* 2014. Semantic frame identification with distributed word representations. In Proceedings of ACL 2014.
14. *Jain, Anil K.* 2010. Data clustering: 50 years beyond K-means. Pattern recognition letters, Vol. 31 (8), pp. 651–666.
15. *Kashkin, Egor, and Olga Lyashevskaya.* Semanticheskie roli i set' konstrukcij v sisteme FrameBank [Semantic roles and construction net in Russian FrameBank]. In: Computational linguistics and intellectual technologies. Proceedings of International Workshop Dialogue'2013. Vol. 12 (19), 2013. Moscow: RGGU. Pp. 325–343.
16. *Korhonen, Anna.* 2002. Subcategorization Acquisition. Ph.D. thesis, University of Cambridge, Computer Laboratory. Technical Report UCAM-CL-TR-530.
17. *Kuznetsov, Ilya.* Semantic Role Labeling for Russian language based on Russian FrameBank // AIST-2015. CCIS, Springer (forthcoming).
18. *Kuznetsova, Era V.* (ed.). 1989. Lexico-semantic groups of Russian verbs [Leksiko-semanticheskie gruppy russkikh glagolov]. Irkutsk.
19. *Kuznetsova, Julia, and Olga Lyashevskaya.* Konstrukcii i transformacii [Constructions and transformations]. Electronic publication: Slovo i Jazyk, 2–4 February 2010, Moscow, Russia. IPPI RAN.
20. *Lang, Joel and Mirella Lapata.* 2011. Unsupervised semantic role induction with graph partitioning, Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pp. 1320–1331.
21. *Lang, Joel, and Mirella Lapata.* 2014. Similarity-driven semantic role induction via graph partitioning. Computational Linguistics 40 (3). Pp. 633–669.
22. *Langfelder, Peter, and Steve Horvath.* 2012. Fast R functions for robust correlations and hierarchical clustering. Journal of statistical software, Vol. 46 (11).
23. *Lapata, Maria.* 1999. Acquiring lexical generalizations from corpora: A case study for diathesis alternations. In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics, College Park, MD. Pp. 397–404.
24. *Lapata, Mirella, and Chris Brew.* 2004. Verb class disambiguation using informative priors. Computational Linguistics, Vol. 30 (1). Pp. 45–73.
25. *Lenci, Alessandro.* 2014. Carving Verb Classes from Corpora. In Simone, Raffaele, Francesca Masini (eds.), Word Classes: Nature, typology and representations. Current Issues in Linguistic Theory 332. John Benjamins, Amsterdam, Philadelphia. Pp. 17–36. http://sesia.humnet.unipi.it/lexit/papers/lenciWordClasses.pdf
26. *Levin, Beth.* 1993. English Verb Classes and Alternations. University of Chicago Press, Chicago.

27. *Lyashevskaya, Olga, Kashkin, Egor.* Evaluation of frame-semantic role labeling in a case-marking language. In: Computational linguistics and intellectual technologies. Proceedings of International Workshop Dialogue'2014. Vol. 20. Pp. 362–378.
28. *Lyashevskaya, Olga and Julia Kuznetsova.* Russkij FrameNet: k zadache sozdanija korpusnogo slovarja konstrukcij [Russian FrameNet: towards a corpus-based dictionary of constructions]. In: Computational linguistics and intellectual technologies. Proceedings of International Workshop Dialogue'2009. Vol. 8 (15), 2009. Moscow: RGGU. Pp. 306–312.
29. *Palmer, Martha.* 2009. Semlink: Linking PropBank, VerbNet and FrameNet. In Proceedings of the Generative Lexicon Conference. Sept. 2009. GenLex: Pisa, Italy.
30. *Pustejovsky, James.* 1995. The Generative Lexicon, Cambridge, Mass.: MIT Press.
31. *Schulte im Walde, Sabine.* 2006. Experiments on the Automatic Induction of German Semantic Verb Classes. Computational Linguistics 32:2. Pp. 159–194.
32. *Schulte im Walde, Sabine.* 2009. The Induction of Verb Frames and Verb Classes from Corpora. Corpus Linguistics. An International Handbook ed. by Anke Lüdeling & Merja Kytö. Berlin: Mouton de Gruyter. Pp. 952–972.
33. *Shalyapina, Zoya M.* 2001. Co-occurence valencies as a universal tool for description of natural language syntagmatics [Strukturnye valentnosti kak universal'nyi instrument opisaniya yazykovoi sochetaemosti]. Moscow Journal of Linguistics, 2001, Vol. 5, № 2. Pp. 35–84.
34. *Shvedova, Natal'ya Ju.* 1998–2007. Russian semantic dictionary [Russkiy semanticheskiy slovar']. Moscow.
35. *Smirnov I. V., Shelmanov A. O., Kuznetsova E. S. and Hramoin I. V.* 2014. Semantic-syntactic analysis of natural languages. Part II. Method for semantic-syntactic analysis of texts [Semantiko-sintaksicheskij analiz estestvennyh jazykov II. Metod semantiko-sintaksicheskogo analiza tekstov]. Artificial Intelligence and Decision Making [Iskusstvennyj intellekt i prinjatie reshenij], Vol. 1, pp. 95–108.
36. *Täckström, Oscar, Kuzman Ganchev, Dipanjan Das.* 2015. Efficient Inference and Structured Learning for Semantic Role Labeling. Transactions of the Association for Computational Linguistics, Vol. 3. Pp. 29–41.
37. *Wierzbicka, Anna.* 1983. Genry mowy. In T. Dobrzyńska, E. Janus (eds.) Tekst i zdanie: Zbiór studiów. Wrocław: Ossolineum. Pp. 125–137.

EXERCISE MAKER: АВТОМАТИЧЕСКОЕ СОЗДАНИЕ ЯЗЫКОВЫХ УПРАЖНЕНИЙ

Малафеев А. Ю. (aumalafeev@hse.ru)

Национальный исследовательский университет
«Высшая школа экономики», Нижний Новгород, Россия

Ключевые слова: автоматическое создание языковых упражнений, автоматическая обработка текста, обучение иностранному языку с помощью компьютера, лексико-грамматические упражнения, анализ сложности текста, английский язык как иностранный

EXERCISE MAKER: AUTOMATIC LANGUAGE EXERCISE GENERATION

Malafeev A. Yu. (aumalafeev@hse.ru)

National Research University Higher School of Economics,
Nizhny Novgorod, Russia

Current trends in education, namely blended learning and computer-assisted language learning, underlie the growing interest to the task of automatically generating language exercises. Such automatic systems are especially in demand given the variability in language learning. Despite the abundance of resources for language learning, there is often a lack of specific exercises targeting a particular group of learners or ESP course. This paper gives an overview of a computer system called Exercise Maker that is aimed at flexible and versatile language exercise generation. The system supports seven exercise types, which can be generated from arbitrary passages written in English. Being able to tailor educational material to learners' interests is known to boost motivation in learners (Heilman et al., 2010). An important feature of the system is the automatic ranking of the source passages according to their complexity/readability. As shown by expert evaluation, the automatically generated exercises are of high quality: the gap precision is about 97–98%, while the overall exercise acceptance rate varies from 90% to 97.5%. Exercise Maker is freely available for educational and research purposes.

Key words: language exercise generation, natural language processing, computer-assisted language learning, lexico-grammatical exercises, text readability, English as a foreign language (EFL)

1. Introduction

Although commonly used, language tests and exercises are expensive to create manually. To address this issue, several systems for automatic language exercise generation have been developed in the past two decades. The exercises that can be automatically generated by these systems vary greatly in terms of the target aspects of vocabulary and grammar, supported languages, flexibility and effectiveness (see the next section for more details). From the pedagogic perspective, using automatically generated content seems especially relevant considering the modern educational trends, namely blended learning (Graham, 2006) and computer-assisted language learning (Levy, 1997).

In this paper, we give an overview of a computer program called "Exercise Maker" that we developed for automatic generation of lexical and grammatical exercises from arbitrary passages written in English. The program features seven different exercise types that can be customized to accommodate various learning needs. Importantly, Exercise Maker allows the user to adjust the difficulty: even if the same source passage is used, the exercises can be more challenging or less demanding. Currently, a Japanese version of the program is also under development.

The paper is structured as follows: Section 2 presents an overview of existing systems for automatic language exercise generation; Section 3 describes our solution; Section 4 presents some results of the evaluation of the effectiveness of the system; in the last section, we make conclusions and outline some possible ways of improving Exercise Maker.

2. Related Work

Some recent research has been conducted with a view to facilitating exercise creation. Among the more general solutions are multi-domain exercise or test generation systems, e.g. (Almeida et al., 2013; Mitkov et al., 2006; Sonntag, 2009) exercise generation systems usually work with answers of simple types (e.g. multiple-choice, Boolean, integer, or file comparison, as well as authoring tools, e.g. Hot Potatoes (http://hotpot.uvic.ca/), MaxAuthor (http://cali.arizona.edu/docs/wmaxa/) and others.

There are also several systems that are designed for generating exercises of one or more specific types to aid learners of the supported language(s). These are very different not only in the types of exercises they are able to generate, but also in terms of supported languages, type of input/output, external dependencies and whether the system is freely available. See Tab. 1 for details on these differences and a comparison of Exercise Maker with other systems. This is by no means a complete list of exercise generating systems, but rather some of the better-known and often cited ones. We believe that this is sufficient to demonstrate the prevalent trends in exercise generation and how Exercise Maker attempts to 'fill' some of the 'gaps'.

Tab. 1. Comparison of systems

No.	System	Languages	Input	Output	Exercise type(s)	External dependencies	Freely available
1	(Aldabe et al., 2006)	Basque	corpora	sentences	**fill-in-the-blank**, **word formation**, multiple choice, **error correction**	corpora; morpho-syntactic and syntactic parsers, phrase chunker	?
2	(Antonsen et al., 2013)	two Saami languages	lexicon and syntactic rules	sentences	**morphological transformation**	**none**	**yes**
3	(Bick, 2005)	**English** and 6 other	corpora	sentences	**open cloze**, **morphological transformation**	corpora	**yes**
4	(Brown et al., 2005)	**English**	words	questions	definition, synonym, antonym, hypernym, hyponym, and cloze questions (multiple choice or **wordbank**)	WordNet; external word frequency database	yes?
5	(Burstein and Marcu, 2005)	Arabic → **English**	corpora	sentences	translation	corpora; Arabic-to-English machine translation system	no
6	(Dickinson and Herring, 2008)	Russian	lexicon and syntactic rules	sentences	**morphological transformation**, **error correction**	none	no
7	(Gates, 2008)	**English**	texts from a corpus	questions	reading comprehension questions (factoid)	corpus; syntactic parser, lemmatizer, named entity extractor, semantic argument extractor, WordNet, parse tree transformer	no
8	(Goto et al., 2010)	**English**	**arbitrary texts**	questions	multiple choice	POS-tagger, web search	no
9	(Heilman and Eskenazi, 2007)	**English**	thesaurus	questions	finding related words	dependency parser, corpus	yes?
10	(Hoshino and Nakagawa, 2005)	**English**	**arbitrary texts**	sentences	multiple choice	WordNet	no
11	(Knoop and Wilske, 2013)	**Enslish**	**arbitrary texts**	sentences	multiple choice	WordNet	no
12	(Meurers et al., 2010)	**English**	arbitrary web-pages	**text**	**morphological transformation**, multiple choice, **open cloze**	external NLP framework; a separate external POS-tagger and constraint grammar rules; lexical database	yes
13	(Perez-Beltrachini et al., 2012)	French	lexicon and syntactic rules	sentences	shuffle questions, **open cloze**	grammar traverser	no?
14	(Sumita et al., 2005)	**English**	corpora	sentences	multiple choice	corpora; web search	no
15	Exercise Maker	**English**	**arbitrary texts**	**text**	fill in missing words (no blanks), **open cloze**, word formation, **wordbank**, **morphological transformation** (verb forms), text fragments, **error correction**	**none**	**yes**

To summarize, Exercise Maker is significantly different from most other systems: although it supports English only, indeed a very popular language, it generates exercises from arbitrary passages, which is a feature of only three other systems. Moreover, the output is also text, i.e. the exercises are not sets of separate, unrelated sentences, like in most other systems, but the same passages as input, with some modifications (e.g. gapped words, artificial 'errors', etc.). This 'context-rich' format is very similar to the one used in Cambridge English certificate exams, such as FCE, CAE, CPE, and BEC (e.g. see Cambridge English: Advanced Handbook for Teachers, 2012), which are very well-known and well-established English language tests (Chalhoub-Deville and Turner, 2000). In addition, the same format is used in the Russian State Exam (RSE) in English. But perhaps most importantly, being able to use any passages in English (rather than corpora or grammars and lexicons) means an opportunity to tailor educational material to learners' interests, which is known to boost learner motivation (Heilman et al., 2010).

Furthermore, some of the exercise types in Exercise Maker are not supported by other systems, namely filling in missing words (no gaps), word formation, and text fragments. Yet these types of exercises are commonly used in EFL, and some of them are included in FCE, CAE and CPE (word formation) and the RSE (word formation and text fragments). The difficulty of the exercises can be tweaked, which, although not shown in Tab. 1, is a very rare feature. Another important difference is that Exercise Maker is fully self-contained, which means that it can be more easily extended to resource-poor languages. Lastly, our system is freely available to anyone and, therefore, can be used not only for teaching and learning English, but also for research purposes, e.g. for comparison with other exercise-generating systems.

The next section will discuss the methods used for generating lexico-grammatical exercises in the Exercise Maker system.

3. Automatic Exercise Generation

Exercise Maker supports seven exercise types, which are listed in Tab. 2 with some additional information. This includes which exams, if any, use this type of task, as well as short descriptions and examples for each supported activity. The examples are generated by Exercise Maker using an input passage adapted from a Wikipedia article (http://en.wikipedia.org/wiki/Aron_Ralston).

Our system is implemented in Python and uses the standard libraries only. The generation method used is decision trees with manually written rules, although the exact algorithms vary depending on the exercise type. The rules often involve consulting a set of linguistic resources, specifically compiled by the author (manually and semi-automatically) for exercise generation. The linguistic resources are:

1. Two lists of 2274 and 10084 most common English word forms (including proper nouns), based on a free film-subtitle-based frequency list (https://invokeit.wordpress.com/frequency-word-lists/).
2. A list of 11805 word forms used in the word formation exercise heavily based on the BNC lists (http://simple.wiktionary.org/wiki/Wiktionary:BNC_spoken_freq).

3. A list of rules for making realistic spelling/lexical/grammar errors (795 words). The spelling part is based on the Wikipedia list of common misspellings (http://en.wikipedia.org/wiki/Wikipedia:Lists_of_common_misspellings), while the lexical and grammar error rules were compiled manually.
4. Three ordered lists of 139 words each for generating open cloze tests emulating specific Cambridge exam levels (FCE, CAE or CPE), based on an empirical study of the mentioned exams.
5. A list of 91 adverbs used in the verb forms exercise.
6. A list of 13540 verb forms and an additional short list of auxiliary forms, both used in the verb forms exercise. The lists were extracted from the Spelling Checker Oriented Word List (http://wordlist.sourceforge.net/).
7. A few manually written shorter lists of articles, conjunctions, prepositions, pronouns, etc.

Tab. 2. Exercise types supported by Exercise Maker

No.	Exercise	Exams	Description	Example(s)	Answer(s)
1	Word formation	FCE, CAE, CPE, RSE	Fill in blanks with derivatives of the words in parentheses.	...but the tools he had available were (6)_____ (sufficient) to do so.	insufficient
2	Error correction	BEC[1]	Correct spelling/lexical/grammar errors in the text.	Ralston had not informed nobody of his hiking plans <...> thus no one would searching for him <…> the dehydrated and deliriouse Ralston	had not informed anybody; no one would search/ be searching for him; delirious
3	Open cloze	FCE, CAE, CPE, BEC	Fill in blanks with suitable words (no candidate answers given). Sometimes, there are two or more correct answers.	When he ran (12)_____ of food and water on the fifth day…	out
4	Wordbank	none[2]	Fill in blanks with suitable words given a full list of answer choices (no distractors; each word is used only once).	(approximately, available, <…>, just, suspended) ...a (2)_____ boulder he was climbing down became dislodged…	suspended
5	Missing words (articles or prepositions)	none	Insert prepositions (another subtype: articles) where appropriate.	Ralston had not informed anybody his hiking plans, thus no one would be searching him.	Ralston had not informed anybody of his hiking plans, thus no one would be searching for him.
6	Text fragments	RSE	Insert missing text fragments (all answer options are listed).	After three days of trying to lift (6)_____, the dehydrated and delirious Ralston	d) and break the boulder
7	Verb forms	RSE	Use the appropriate verb form to fill each of the gaps.	While he (1)_____ (descend) a slot canyon, a suspended boulder…	was descending

[1] A somewhat similar task, but with only one error type—extra words.
[2] However, most exams use a somewhat similar test, the multiple choice.

It should be noted that the generation process does not consist in merely looking up words in the mentioned lists. For each type of exercise, there are rules that take into account such factors as capitalization, spelling features, punctuation, word length, distance to other gaps, word context, sentence boundaries, and others. Some rules may be quite complex. For example, dictionary look-ups do not suffice in the missing words (prepositions) exercise, because many words in English are ambiguous with respect to their part of speech. Thus, to determine that *to* is a preposition rather than a particle in a given context (both are common cases), the system checks if the next word is a determiner, or is capitalized, or contains a digit, or is the beginning of a new sentence, or it is longer than five characters and ends in *-ing*.

Another example of using rules beyond dictionary look-ups is making one common error in the error correction exercise. The error is misspelling adjectives ending in *-ous* and *-ful*, such as *furious* and *powerful*, to make these **furiouse* and **powerfull*. It would be difficult to list all possible adjectives that can be modified in this way, so the system uses the following simple rule: if the word is not all caps (avoids clashes with abbreviations) and ends in *-ous* or *-ful*, replace the ending with *-ouse* or *-full*, correspondingly.

These are merely some examples of the rules used in exercise generation; the size constraints do not permit listing all the rules.

The preprocessing step, performed once for each source passage, includes segmenting the input text into words, sentences and paragraphs, and analyzing the readability of the source. Text readability has a number of formal, quantitative characteristics, such as the average number of syllables in words and the average sentence length (Kincaid et al., 1975), or word frequency with respect to either a reference list (Chall, 1995) or corpus data (Stenner, 1996). Admittedly, these measures are error-prone and may be inaccurate at times, but they are still highly useful for approximating source text complexity. In Exercise Maker, the input passages are ranked according to their complexity. The latter is an important feature, as it helps the teacher to select materials appropriate for the ability level of particular learners, as, obviously, the readability of the source passage strongly correlates with the difficulty of the resulting exercises.

After experimenting with some variables, we chose two of them as the main proxy for text readability, namely the average sentence length and word frequency information. These two factors have traditionally been considered as the most closely correlated with text readability (Klare, 1968; Chall, 1995)3,15]]}},"label":"page"},{"id":315,"uris":["http://zotero.org/users/1547774/items/24IK5VWR"],"uri":["http://zotero.org/users/1547774/items/24IK5VWR"],"itemData":{"id":315,"type":"book","title":"Readability revisited: The new Dale-Chall readability formula","publisher":"Brookline Books Cambridge, MA","volume":"118","source":"Google Scholar","shortTitle":"Readability revisited","author":[{"family":"Chall","given":"Jeanne Sternlicht"}],"issued":{"date-parts":[["1995"]]}},"label":"page"}],"schema":"https://github.com/citation-style-language/schema/raw/master/csl-citation.json"} . While the first is very easy to calculate, the second may be approached in various ways. Similarly to (Chall, 1995), we use reference lists to approximate word frequency. Specifically, we have two reference lists of 2,274 and 10,084 most common English word forms, including proper nouns; these represent two levels of word frequency. If a word is in the first list, which means that it is very commonly used in English, it is also a member of the second, larger list.

Common, but less frequent words are those that are in the second list and not in the first. If a word is not a member of either of the lists, it is considered an 'unknown' word. Readability is thus determined by the proportion of 'unknown' words in a text, the proportion of words that are not in the first list, and the average sentence length. These three factors have equal weight in our simple complexity model.

Importantly, Exercise Maker goes beyond readability in adjusting the difficulty of exercises. For almost every type of exercise (except missing words), the system generates several subtypes with varying settings that affect the difficulty. These settings are:

- number of gaps in the exercise;
- target language material, i.e. the words in the text that are gapped (for the open cloze and verb forms exercises);
- length of the gaps (for the fragments exercise).

Regarding target words, it might be necessary to clarify that the open cloze exercises are based on five different lists of target word forms, which is aimed at generating exercises of varying difficulty and at emulating specific Cambridge exam task types: FCE, CAE and CPE. As for the verb form exercises, these come in two varieties: gapping 'simple', one-word verb forms and more complex, multiword verb forms.

In the next section, we describe an experiment conducted to evaluate the performance of our system.

4. Evaluation

Although TEFL experts in several educational establishments have successfully used our system, it is necessary to present here a formal evaluation of Exercise Maker. Earlier, we evaluated specific exercise types such as the open cloze (Malafeev, 2014) and got interesting results. In particular, two groups of TEFL experts (17 and 16 people) found it considerably difficult to tell the difference between activities generated by Exercise Maker and tests authored by Cambridge professionals.

For this publication, we conducted a specific evaluation session covering all seven exercise types. Two independent TEFL experts, both non-native speakers of English, who had not taken any part in developing Exercise Maker, participated in the evaluation. We downloaded five abridged and simplified news articles from a popular website for EFL teachers and learners, breakingnewsenglish.com (see Tab. 3). We had not read or otherwise used these articles prior to the evaluation experiment.

Tab. 3. News articles used for evaluation

No.	Title	Date	Word count	Readability[1]
1	Japanese government to play matchmaker	17th March, 2015	243	10.7
2	BBC Top Gear star punches producer	14th March 2015	237	7.1
3	Sportswear maker accused of sexism	11th March, 2015	231	12.1
4	China tops US at box office for first time	5th March, 2015	233	7.7
5	Cut music to an hour a day	2nd March, 2015	248	12.2

[1] For readability, we used the Automated Readability Index (Kincaid et al., 1975), calculated using the Edit Central online service (http://www.editcentral.com/gwt1/EditCentral.html).

The five articles were used as input to generate 40 exercises with our system, eight from each text. Although the number of exercise types supported by the system is seven, we chose to use two different subtypes of the missing words exercise, namely articles and prepositions.

The experts had to perform two kinds of assessment:

- evaluate all gaps in all exercises and determine which of the gaps are valid, i.e. potentially useful in teaching or testing, and which are not (evaluating precision only is a widely accepted practice in automatically-generated exercise evaluation);
- assign to each exercise an overall score from 1 to 4, meaning:
  1 – the exercise cannot be used;
  2 – the exercise can be used only after making substantial alterations;
  3 – the exercise can be used, but it requires some minor alterations;
  4 – the exercise can be used as is.

In the latter form of assessment, the scores of 3 or 4 would mean that the exercise is 'acceptable', and the lower scores would mean that it is not.

The experts were supplied with detailed instructions written in Russian on the evaluation procedure. The evaluation took about three hours (expert 1) and five hours (expert 2). We believe that, given that both the articles and our system are freely available for download, and the assessment guidelines are available on request, our experiment can be easily reproduced, although, of course, with different experts. The results of the evaluation are presented in Tab. 4, Tab. 5 and Tab. 6.

Tab. 4. Evaluation results, validity of gaps (precision)

	Evaluation	Total gaps	Expert 1, n	Expert 1	Expert 2, n	Expert 2
Exercises	Articles	110	109	99,09%	110	100,00%
	Derivatives	60	52	86,67%	57	95,00%
	Errors	131	131	100,00%	129	98,47%
	Fragments	30	30	100,00%	30	100,00%
	Open cloze	144	142	98,61%	144	100,00%
	Prepositions	137	126	91,97%	131	95,62%
	Verb forms	73	73	100,00%	69	94,52%
	Wordbank	100	97	97,00%	100	100,00%
Texts	1	153	145	94,77%	151	98,69%
	2	171	165	96,49%	168	98,25%
	3	153	147	96,08%	146	95,42%
	4	144	142	98,61%	142	98,61%
	5	164	161	98,17%	163	99,39%
Total	micro	785	760	96,82%	770	98,09%
	macro, exercises			96,67%		97,95%
	macro, texts			96,92%		97,80%

Tab. 5. Evaluation results, accepted exercises

	Evaluation	Total exercises	Accepted by expert 1		Accepted by expert 2	
			n	%	n	%
Exercises	Articles	5	5	100,00%	5	100,00%
	Derivatives	5	4	80,00%	4	80,00%
	Errors	5	5	100,00%	5	100,00%
	Fragments	5	4	80,00%	5	100,00%
	Open cloze	5	5	100,00%	5	100,00%
	Prepositions	5	3	60,00%	5	100,00%
	Verb forms	5	5	100,00%	5	100,00%
	Wordbank	5	5	100,00%	5	100,00%
Texts	1	8	6	75,00%	8	100,00%
	2	8	7	87,50%	8	100,00%
	3	8	8	100,00%	7	87,50%
	4	8	7	87,50%	8	100,00%
	5	8	8	100,00%	8	100,00%
Total		40	36	**90,00%**	39	**97,50%**

Tab. 6. Scores assigned by the experts

Score	Expert 1		Expert 2	
	n	%	n	%
1	0	0,00%	0	0,00%
2	4	10,00%	1	2,50%
3	22	55,00%	8	20,00%
4	14	35,00%	31	77,50%

As can be seen from the tables, the gap precision is about 97–98%, which is very high. The acceptance rate varies significantly, from 90% (expert 1) to 97.5% (expert 2). This difference can probably be explained by the fact that the borderline between "substantial alterations" and "minor alterations" is not well-defined and depends on the subjective judgment, even with assessment guidelines. Besides, as commented by the first expert, while individual gaps seemed valid, the combination of these did not always produce a good exercise. Indeed, different TEFL professionals might have varying opinions about what exactly constitutes a good language exercise. Still, we believe that even the lower, 90% acceptance rate is a very good result for automatic language exercise generation.

The next section will draw conclusions and outline some possible directions for future work.

5. Conclusion

This paper presents an overview of our language exercise generation system, Exercise Maker. With it, a variety of lexical and grammatical exercises can be automatically generated from arbitrary passages written in English. The source passages are ranked according to their readability to help the user choose appropriate material. The seven types of supported exercises can be further customized to accommodate various learning needs. Besides, Exercise Maker allows the user to adjust the difficulty, even if the same source passage is used. As shown in the evaluation section, the exercises generated are perceived by TEFL experts as quite useful.

The most promising directions of future work are the following:

- support for other languages;
- new exercise types, such as multiple choice;
- further improving exercise quality, possibly with statistical methods and machine learning.

Acknowledgements

The author thanks the three anonymous reviewers for valuable comments and suggestions.

References

1. *Aldabe, I., Lacalle, M. L. de, Maritxalar, M., Martinez, E., and Uria, L.* (2006). ArikIturri: An Automatic Question Generator Based on Corpora and NLP Techniques. In Intelligent Tutoring Systems, M. Ikeda, K. D. Ashley, and T.-W. Chan, eds. (Springer Berlin Heidelberg), pp. 584–594.
2. *Almeida, J. J., Araujo, I., Brito, I., Carvalho, N., Machado, G. J., Pereira, R. M. S., and Smirnov, G.* (2013). PASSAROLA: High-Order Exercise Generation System. In 2013 8th Iberian Conference on Information Systems and Technologies (CISTI), pp. 1–5.
3. *Antonsen, L., Johnson, R., Trondsterud, T., and Uibo, H.* (2013). Generating Modular Grammar Exercises with Finite-State Transducers. Proceedings of the Second Workshop on NLP for Computer-Assisted Language Learning at NODALIDA 2013 27–38.
4. *Bick, E.* (2005). Live Use of Corpus Data and Corpus Annotation Tools in CALL: Some New Developments in VISL. Nordic Language Technology, AArbog for Nordisk Sprogteknologisk Forskningsprogram 171–185.
5. *Brown, J. C., Frishkoff, G. A., and Eskenazi, M.* (2005). Automatic Question Generation for Vocabulary Assessment. In Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing, (Stroudsburg, PA, USA: Association for Computational Linguistics), pp. 819–826.

6. *Burstein, J., and Marcu, D.* (2005). Translation Exercise Assistant: Automated Generation of Translation Exercises for native-Arabic Speakers Learning English. In Proceedings of HLT/EMNLP on Interactive Demonstrations, (Stroudsburg, PA, USA: Association for Computational Linguistics), pp. 16–17.
7. *Chalhoub-Deville, M., and Turner, C. E.* (2000). What to Look for in ESL Admission Tests: Cambridge Certificate Exams, IELTS, and TOEFL. System 28, 523–539.
8. *Chall, J. S.* (1995). Readability revisited: The new Dale-Chall readability formula (Brookline Books Cambridge, MA).
9. *Dickinson, M., and Herring, J.* (2008). Developing Online ICALL Exercises for Russian. In Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications, (Stroudsburg, PA, USA: Association for Computational Linguistics), pp. 1–9.
10. *Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., and Willingham, D. T.* (2013). Improving Students' Learning With Effective Learning Techniques Promising Directions From Cognitive and Educational Psychology. Psychological Science in the Public Interest 14, 4–58.
11. *Gates, D. M.* (2008). Automatically Generating Reading Comprehension Look-Back Strategy: Questions from Expository Texts (Pittsburgh: Carnegie Mellon University).
12. *Goto, T., Kojiri, T., Watanabe, T., Iwata, T., and Yamada, T.* (2010). Automatic Generation System of Multiple-Choice Cloze Questions and its Evaluation. Knowledge Management & E-Learning: An International Journal (KM&EL) 2, 210–224.
13. *Graham, C. R.* (2006). Blended learning systems. CJ Bonk & CR Graham, The Handbook of Blended Learning: Global Perspectives, Local Designs. Pfeiffer.
14. *Heilman, M., and Eskenazi, M.* (2007). Application of Automatic Thesaurus Extraction for Computer Generation of Vocabulary Questions. In Proceedings of the SLaTE Workshop on Speech and Language Technology in Education, pp. 65–68.
15. *Heilman, M., Collins-Thompson, K., Callan, J., Eskenazi, M., Juffs, A., and Wilson, L.* (2010). Personalization of Reading Passages Improves Vocabulary Acquisition. International Journal of Artificial Intelligence in Education 20, 73–98.
16. *Hoshino, A., and Nakagawa, H.* (2005). WebExperimenter for Multiple-choice Question Generation. In Proceedings of HLT/EMNLP on Interactive Demonstrations, (Stroudsburg, PA, USA: Association for Computational Linguistics), pp. 18–19.
17. *Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., and Chissom, B. S.* (1975). Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel (DTIC Document).
18. *Klare, G. R.* (1968). The Role of Word Frequency in Readability. Elementary English 12–22.
19. *Knoop, S., and Wilske, S.* (2013). WordGap—Automatic Generation of Gap-Filling Vocabulary Exercises for Mobile Learning. Proceedings of the Second Workshop on NLP for Computer-Assisted Language Learning at NODALIDA 2013 39–47.
20. *Levy, M.* (1997). Computer-Assisted Language Learning: Context and Conceptualization. (ERIC).

21. *Malafeev, A.* (2014). Automatic Generation of Text-Based Open Cloze Exercises. In Analysis of Images, Social Networks and Texts, D. I. Ignatov, M. Y. Khachay, A. Panchenko, N. Konstantinova, and R. E. Yavorsky, eds. (Cham: Springer International Publishing), pp. 140–151.
22. *Meurers, D., Ziai, R., Amaral, L., Boyd, A., Dimitrov, A., Metcalf, V., and Ott, N.* (2010). Enhancing Authentic Web Pages for Language Learners. In Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications, (Stroudsburg, PA, USA: Association for Computational Linguistics), pp. 10–18.
23. *Mitkov, R., An Ha, L., and Karamanis, N.* (2006). A Computer-Aided Environment for Generating Multiple-Choice Test Items. Natural Language Engineering 12, 177–194.
24. *Perez-Beltrachini, L., Gardent, C., and Kruszewski, G.* (2012). Generating Grammar Exercises. In Proceedings of the Seventh Workshop on Building Educational Applications Using NLP, (Stroudsburg, PA, USA: Association for Computational Linguistics), pp. 147–156.
25. *Sonntag, M.* (2009). Exercise Generation by Group Models for Autonomous Web-Based Learning. In 35th Euromicro Conference on Software Engineering and Advanced Applications, 2009. SEAA '09, pp. 57–63.
26. *Stenner, A. J.* (1996). Measuring Reading Comprehension with the Lexile Framework. Fourth North American Conference on Adolescent/Adult Literacy, Washington DC.
27. *Sumita, E., Sugaya, F., and Yamamoto, S.* (2005). Measuring Non-native Speakers' Proficiency of English by Using a Test with Automatically-generated Fill-in-the-blank Questions. In Proceedings of the Second Workshop on Building Educational Applications Using NLP, (Stroudsburg, PA, USA: Association for Computational Linguistics), pp. 61–68.
28. *Cambridge English* (2012): Advanced Handbook for Teachers (Cambridge ESOL).

МЕТАЯЗЫКОВОЙ ПОРТРЕТ МОДНЫХ СЛОВ

Мустайоки А. (arto.mustajoki@helsinki.fi)
Хельсинкский университет, Хельсинки, Финляндия

Вепрева И. Т. (irina_vepreva@mail.ru)
Уральский федеральный университет,
Екатеринбург, Россия

В статье проводится корпусное исследование сущностных характеристик модного слова. Поисковым инструментом для выявления класса искомых единиц являются высказывания, включающие метаоператоры *модное слово, как модно говорить*. В основе интерпретации материала лежит теоретическая модель моды, разработанная А.Б. Гофманом. В работе определяются контекстные маркеры, манифестирующие атрибутивные признаки модного объекта: «современность», «универсальность», «демонстративность» и «игра».

На основе корпуса метаязыковых оценок выделяются три класса модных слов, которые обыденным сознанием объединены двумя атрибутивными ценностями моды — параметрами современности и универсальности. К ним относятся актуализированные высокочастотные слова (1); новые слова, называющие новую реалию (2); слова, являющиеся новым обозначением известного явления (3). Только последний класс слов отвечает всем предъявляемым критериям моды, реализуя игровое начало — эстетическую потребность говорящего к обновлению речи.

К маркерам демонстративности были отнесены контекстные указатели необычности формальной стороны модного слова. Внешней привлекательностью обладают, прежде всего, иноязычные неологизмы.

Класс модных лексем можно представить как полевую структуру с ядром слов, удовлетворяющих всем предъявляемым ценностным критериям модного объекта, и периферией.

Ключевые слова: модное слово, метаязыковое высказывание, атрибутивные ценности моды, современность, универсальность, демонстративность, игра

METALINGUISTIC PORTRAIT OF FASHIONABLE WORDS

Mustajoki A. (arto.mustajoki@helsinki.fi)
University of Helsinki, Helsinki, Finland

Vepreva I. T. (irina_vepreva@mail.ru)
Ural Federation University, Ekaterinburg, Russia

The paper is based on a corpus study on the essential characteristics of fashionable word. The retrieval system for the identification of basic units consists of the utterances, containing metaoperators *modnoe slovo* (fashionable or trendy word) and *kak modno govorit'* (how to speak in a fashionably way). The theoretical model of fashion, worked out by A.B. Gofman, served as the basis for the interpretation of the findings. The work distinguishes context markers, manifesting attributive characteristics of the fashionable object, *modernity, universality, demonstrativeness* and *play*.

Based on the metalinguistic valuations corpus there were distinguished three classes of fashionable words, associated by commonplace consciousness with two attributive fashionable values—modernity and universality. These words include 1) the new words, naming the new reality; 2) the words, referring to the new naming of the known reality; 3) the only class of words meeting all the requirements of fashion, realizing play activity or speaker's aesthetic need to renew his speech.

The demonstrativeness markers are distinguished by the context indicators of the unusualness of the fashionable word form. In the first instance, the foreign neologisms demonstrate external attractiveness.

The class of the fashionable lexical items can be presented as a field structure with a nexus, having all the required value criteria of a fashionable object, and peripheral layers of various degree, depending on the number of criteria they have.

Key words: fashionable word, buzzword, trend word, metalinguistic utterance, attributive values of fashion, modernity, universality, demonstrativeness, play

1. Введение

Благодаря современным технологиям в поле зрения лингвистов оказались крупные корпуса текстов. Сильной стороной корпусного подхода является возможность изучать те языковые объекты, которые были неподъемны для анализа в режиме ручного сбора материала. К такому роду языковых явлений можно отнести феномен модного слова. Имеющийся в нашем распоряжении материал был собран с помощью системы базы данных Интегрум (www.integrum.ru), включающей большинство выходящих в настоящее время публицистических текстов и содержащий многофункциональную информационную поисковую систему, реализуя достижения корпусной лингвистики и Интернета.

Основанием для выборки материала, поисковым инструментом для выявления класса лексических единиц, к которым приложима характеристика модных, являются высказывания, включающие прежде всего метаоператоры *модное слово, как модно говорить*. Безусловно, круг метаоператоров, характеризующих модное слово, гораздо шире. К ним относится целый ряд как клишированных, так и индивидуально авторских косвенных метамаркеров: *пущено в ход слово; теперь в моде слово; популярное словечко; слово употребляется часто; слово у всех на слуху; надоела слово, но никуда от него не денешься;*

как принято говорить; как любят говорить и др. В рамках данной работы мы ограничили выборку материала, включив контексты только с прямыми операторами.

Данные высказывания в изоляции от текстов, в которых они употребляются, образуют особого рода дискурс, способный пролить свет на природу интересуемого объекта. Так, корпус рефлексивных высказываний по поводу модных слов 2014–2015 годов формируют, к примеру, такие контексты (о специфике слов, входящих в иллюстративную подборку, разговор пойдет ниже):

(1) *Называться это будет красивым* ***модным словом «импортозамещение»****, то бишь замещение иностранных товаров своими, кровными* (Правда Москвы; 28.08.2014)*;*

(2) *В этом году главное* ***модное слово — «Крымнаш»*** (Собеседник; 24.12.2014)*;*

(3) ***Модное слово санкции*** *стало универсальным паролем, с помощью которого ведомственные лоббисты пытаются получить доступ к бюджетному финансированию* (Новая газета; 08.10.2014)*;*

(4) *Изобретательный русский ум стал мыслить в наиболее лёгком для него направлении: как поскорее сколотить капитал, используя* ***модное слово «инновации»*** *и другие похожие словечки* (АиФ — Москва; 04.03.2015).

(5) *Моя жена в последнее время стала часто употреблять* ***модное слово гаджеты*** (Комсомольская правда; 22.01.2015).

(6) *Как пояснил врио заместителя губернатора Александр Криволапов, тема эта очень актуальная, говоря* ***модным словом, — тренд*** (Курский вестник; 11.06.2014).

Предлагаемый подход к изучению объекта — через метаязыковые показания говорящего — имеет право на свое существование. Точкой отсчета исследовательского интереса к изучению метаязыкового обыденного сознания принято считать дискуссию на конференции 1964 года в Калифорнийском университете UCLA (University of Califorina, Los Angeles), в ходе которой проблема традиционного донаучного знания языка была поставлена на обсуждение в докладе Хенигсвальда и была названа «народной лингвистикой» (folk-linguistics) [Hoenigswald 1966]. Отношение ученых к показаниям обыденного сознания носит неоднозначный характер. С одной стороны, общеизвестно, что лингвистика народная (наивная, обыденная, любительская) несистематична, неполноценна в силу некритичности в отношении собственных продуктов, эмоциональности и пристрастности, мешающих объективной оценке происходящего, с другой стороны, лингвисты сходятся во мнении, что «иногда неискушенные носители языка поражают точностью своих метаязыковых

комментариев — настолько, что „стихийная лингвистика“ оказывается чрезвычайно близка „научной лингвистике“» [Булыгина, Шмелев 2000: 14].

На наш взгляд, метаязыковые комментарии и оценки модного слова, собранные в единый текстовый массив, представляют собой естественно складывающуюся уникальную диагностическую базу данных, важную в методологическом отношении, поскольку данный корпус высказываний проливает свет на лингвосоциопсихологическую сущность феномена языковой моды, явления не собственно языкового характера.

Мода — это, в первую очередь, социально-психологическое явление, которое относится к механизмам социальной регуляции и саморегуляции человеческого поведения, представляя собой «коллективное подражание регулярно появляющимся новинкам» [Барт 1997: 7]. Несмотря на постоянный интерес к моде со стороны культурологов, философов, психологов, социологов, см., например: [Барт 2003, Бодрийяр 2000, Дергунова 2004, Килошенко 2001, Михалева 2010, Орлова 1989, Такер 2003, Ятина 2004 и др.], специальных лингвистических работ, посвященных языковой моде, немного, см.: [Mjos, Moe, Sundet 2014; Tsun-Jui, Shu-Kai, Prevot 2013; Бічай 2003; Новиков 2005, 2012; Титкова 1998; Федорова 2014].

При выявлении параметров модного слова [Мустайоки, Вепрева 2006] мы опирались на теоретическую модель моды, предложенную российским социологом А. Б. Гофманом [Гофман 2000]. По Гофману, любой объект становится модным, когда он начинает обладать модными значениями, выступающими знаками моды. Структурообразующим компонентом моды является набор атрибутивных ценностей, к которым автор относит современность, универсальность, демонстративность и игру [там же: 20–33].

После введения к нашей теме мы готовы сформулировать цель данной работы — с опорой на материалы корпусного ресурса определить сущностные характеристики слова, которые отвечают параметрам модного объекта, выявить специфику обыденного понимания модного слова.

2. Анализ сущностных параметров модного слова

Путем наложения ценностных значений модного объекта на лексемы, которые обыденное языковое сознание определяет как модные, мы выделили коррелятивные признаки модного слова.

Наивный лексиколог, в нашей случае это журналист, к модным относит три достаточно самостоятельных класса слов, которые объединены двумя атрибутивными ценностями моды — параметрами современности и универсальности. Определим данные характеристики модного слова.

2.1. Современность модного слова

Данный параметр предполагает выделение у модного объекта признака темпоральности. Модными единицами называются слова, ставшие актуальными

по разным (внеязыковым или внутриязыковым) причинам. Самым частотным временны́м метаоператором модного слова является наречие «сейчас», типичное метавысказывание — «***Это сейчас*** (сегодня, нынче, теперь) ***модное слово***». Признак актуального настоящего момента реализуется как признак современности модного слова — модным может быть только то, что модно сейчас. Поэтому в метаязыковых контекстах модное слово получает определенную локализацию в языковом и временно́м пространстве. Контексты, которые называют не современные модные слова, всегда содержат привязку к конкретному периоду употребления, когда слово было модным:

(7) *Названы самые* ***модные слова прошедшего телесезона*** (Московская правда / Кнопка; 08.09.2006).

(8) *Но* ***модное слово «тюнинг» появилось у нас где-то в начале 90-х****, я думаю* (Радио Свобода — Программы; 11.06.2008).

(9) ***Проблемы****, если воспользоваться одним из самых* ***модных слов двадцатого века****, возникли позже <…>.* ***К консенсусу*** *— это* ***в тот момент было*** *весьма* ***модное слово*** *— мы так и не пришли* (Нева, Санкт-Петербург; 15.06.2010).

Проиллюстрируем локализацию модного слова в языковом пространстве:

(10) ***В этом сезоне у русской богемы*** *стало* ***модным слово «революция»*** (Радио Свобода, Темы дня, 2005, 22.01).

(11) *Есть* ***в корпоративной среде*** *такое* ***модное слово — тимбилдинг*** (Комсомольская правда; 22.06.2011).

(12) ***В среде интеллектуалов в 2010 году*** *наблюдалось немалое брожение,* «***модернизация***» *мгновенно стала* ***модным словом*** (Неприкосновенный запас; 15.02.2011).

(13) ***Сейчас у политологов*** *появилось новое* ***модное слово «тунисизация»*** (Правда-КПРФ, Москва; 11.02.2011).

Параметр локализации в языковом пространстве представлен достаточно размыто. Социальные группы, в рамках которых употребительно модное слово, выделяются на разных основаниях. Это группы лиц по социальному положению, профессиям, занятиям, возрасту, месту проживания, группы, объединяемые общими интересами, и др. Тем не менее, размытость обыденной параметризации не умаляет значимости метаязыковых наблюдений: модная лексика может доминировать как в узких сферах профессионального общения, так и в общем речевом обиходе. При этом очевидно, что набор модных лексем, обусловленный средой бытования слова, может быть различным.

2.2. Универсальность модного слова

С данным параметром связана такая черта моды, как массовость и экстерриториальность. В языке универсальность коррелирует с признаком частотности употребления слова. Модное слово захватывает максимально широкое социальное пространство за необычно малый промежуток времени. Метаязыковые высказывания передают интуитивное представление носителя языка о резком возрастании частотности употребления: ***многие*** *стали употреблять модное слово; звучит едва ли не* ***на всех континентах; замелькало везде***; ***расхожее*** *модное слово; модное слово* ***все чаще*** *выносится в заголовки; модное слово подхватили* ***все;*** *модное слово* ***захлестнуло всю страну****; модное слово обрушивается на нас* ***отовсюду*** *— с заголовков газет, рекламных транспарантов и новостных сводок* и др.

По нашему мнению, метаязыковые ощущения говорящих о модности слова достаточно четко коррелируют с корпусными данными. Покажем данную соотносительность.

Легче проследить резкий рост частотности слова, если появление лексемы четко зафиксировано. С точностью до конкретной даты может быть отмечено вбрасывание в активный речевой обиход политически ангажированных терминов.

Так, например, случилось со словами ***деноминация*** и ***монетизация***, активное вхождение которых в современный лексикон обусловлено внеязыковыми причинами: проведением с января 1998 года деноминации рубля, о которой было объявлено в августе 1997 года, а также изданием приказа № 122 в августе 2004 года о предстоящей (с января 2005 года) монетизации социальных льгот.

Приведем количественные данные роста частотности данных единиц по годам, а в некоторых случаях и по месяцам (корпус Интегрума).

Деноминация: 1991 г. — *5* (употреблений); 1992 г. — *9*; 1993 г. — *22*; 1994 г. — *56*; 1995 г. — *118*; 1996 г. *228;* 1997 г. (январь — июль) — ***197***; (август — декабрь) — ***5751***; 1998 г. — ***5609***; 1999 г. — *1409*; 2000 г. — *1177;* 2001 г. — *1062*; 2002 г. — *1028*; 2003 г. — *1159*; 2004 г. — *1385*. Метавысказывание *модное слово деноминация* зафиксировано в 1998 году.

Монетизация: 2000 г. — *417*; 2001 г. — *595*; 2002 г. — *504*; 2003 г. — *484*; 2004 г. (январь–июль) — ***724***; (август — декабрь) — ***20 955***; 2005 г. январь — ***24 922***; февраль — ***22 249***; март — *11 998*; апрель — *9020*; май — *4191;* июнь — *4267*; июль — *3624*; август — *3125*; сентябрь — *1962*. Метавысказывание *модное слово монетизация* зафиксировано в 2005 году.

Обратимся к современным модным лексемам.

Модернизация: 1999 г. — *28 455*; 2000 г. — *52 324*; 2001 г. — *85 708;* 2002 г. — *85 024*; 2003 г. — *108 901*; 2004 г. — *130 715*; 2005 г. — *157 531*; 2006 г. — *189 094*; 2007 г. — *212 795*; 2008 г. — *228 118*; 2009 г. — *229 330*; 2010 г. — ***336 666***; 2011 г. — ***360 915***; 2012 г. — ***321 321***; 2013 г. — *288 311*; 2014 г. — *276 120*.

Комиссия по модернизации и технологическому развитию экономики России создана в соответствии с указом президента РФ № 579 от 20 мая 2009 г.

Впервые модным слово названо в 2009 году, пик метавысказываний по поводу модности слова приходится на 2010–2012 годы, в 2014 году в корпусе Интегрума не встретилось ни одного «модного» метаконтекста.

На рисунке 1 рост частотности модных слов *модернизация* и *инновация* демонстрирует, что важнее абсолютной частотности — частотность относительная, а именно, резкий рост употребления самого слова без отношения к количественным показателям других слов.

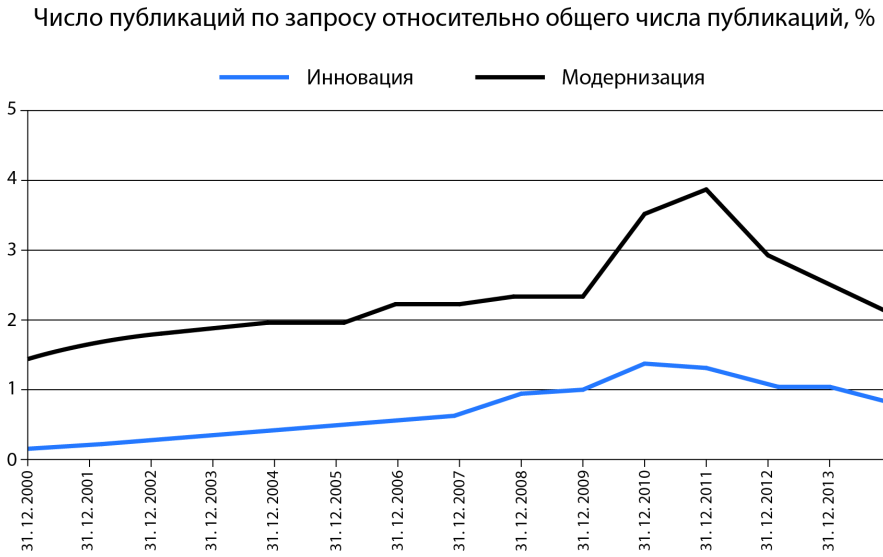


Рис. 1. Рост частотности модных слов *модернизация* и *инновация* (Интегрум)

При анализе корпуса метавысказываний у исследователя может возникнуть ощущение, что журналиста часто подводит языковое чутье, поскольку слово, потерявшее ореол модности, в текстах СМИ по-прежнему называется модным. Данный феномен находит свое объяснение в работах по теории моды, указывающих, что в основе распространения моды лежит психологический механизм подражания, зиммелевский «эффект просачивания»: низшие по социальной лестнице подражают высшим, провинция — центру и т. д.

Продвижение модного слова от центра к провинции можно наблюдать, сравнив встречаемость модного слова в центральной и региональной прессе. Проследим данный процесс на употреблении лексемы ***гламур***. В 1995 году лексема встретилась один раз только в центральной прессе. Дальнейшее распределение единицы выглядит следующим образом: **1996 г.: Ц** (центральная пресса) **— 7** словоупотреблений, **Р** (региональная пресса) — ***1***; 1997 г.: Ц — *8*, Р — *1;* 1998 г.: Ц — *23*, Р — *5*; 1999 г.: Ц — *43*, Р — *15*; 2000 г.: Ц — *90*; Р — *31*; 2001 г.: Ц — *268*, Р — *57*; 2002 г.: Ц — *367*, Р — *132*; 2003 г.: Ц — *740*, Р — *510*; **2004 г. — Ц — *1504*, Р — *1516***; 2005 — Ц — *2027,* Р — *2399;* 2006 — Ц — *3241*; Р. — *4112*.

Впервые модным слово было названо в 2004-м году, когда частотность употребления слова в центральной и региональной прессе практически сравнялась. В 2008-м и 2009-м годах слово продолжало называться модным лишь в региональной прессе, в 2010 году метакомментарий отсутствует во всех СМИ. Отсюда сделаем вывод: метаязыковое комментирование модности происходит в период активного и широкого внедрения модного слова в речевой обиход (говорящий осознает этот этап социализации единицы), а некая метаязыковая погрешность объясняется отставанием провинции в моде. Попутно возникает проблема определения длительности модного цикла в жизни слова, но это особый вопрос, который требует отдельной работы.

Таким образом, актуальность и частотность употребления модного слова для обыденного сознания являются достаточным основанием, чтобы считать единицы модными. Эти признаки явились приоритетными при характеристике модных слов (см. контексты 1–5). Это могут быть «старые», узуальные слова, модными их делает актуализация в современном языковом пространстве (см.: 1, 3–5). Кроме того, модным становится новое для русского языка слово, обозначающее новую реалию (см.: 2).

Широкий взгляд на модные слова носит ненаучный характер и демонстрирует некритичное осмысление языкового объекта обыденным сознанием. Третий класс слов, удовлетворяющий научному подходу к определению модного слова: «модное слово не может обозначать новый денотат, оно всегда является новым обозначением известного явления» [Титкова, 1998: 151, см. об этом также: Розен, 1991: 145–146] — также включается в поле метаязыковой отмеченности носителя языка (см.: 6). Выделению этого класса слов отвечает следующий параметр моды — игра.

2.3. Игровое начало модного слова как смена номинаций

Игре как еще одной внутренней ценности моды присущ эвристический, поисковый характер, который стимулирует смену модных объектов [Гофман 2000: 28–29]. Эстетическая потребность говорящего к обновлению речи реализуется в языке в смене формы языкового знака при тождестве содержания.

Смена номинации постоянно фиксируется в метаязыковом сознании говорящего и может быть отслежена в речевом материале. Языковое сознание выполняет в этом случае регулятивную функцию, сопоставляя «новомодные» и «старомодные» слова. Помещенные в один контекст метаязыкового комментирования, эти единицы становятся временны́ми синонимами, сменяющими друг друга.

Отметим типовые условия реализации данного параметра модного слова:

(14) *Но если раньше это называлось* ***авторским контролем****, теперь —* ***модным словом «аутсорсинг»*** (Красная звезда; 07.04.2012).

(15) *Раньше таких людей называли* ***администраторами****, сейчас —* ***модным словечком «продюсер»*** (Комсомольская правда; 26.11.1999).

В представленных контекстах мы видим типовую организацию высказывания, манифестирующего смену номинаций. В центре высказывания оппозитивная пара единиц, одна из которых получает характеристику *модное слово*. Темпоральная цепочка ***раньше*** (*называли*) — ***теперь, сейчас, нынче*** (*называют*), подчеркивает процессуальный характер феномена моды. Темпоральный уточнитель *раньше* может варьироваться:

(16) *Во время работы с академиком Татаркиным мы нашли подходящее определение, характеризующее нас. Мы назвали это* ***территориально-производственным комплексом*** *— так, как это называлось в СССР. А теперь это называется* ***модным словом «кластер»*** (Эксперт; 13.08.2012).

(17) *Прабабушкин* ***палисадник*** *называется сегодня* ***модным словом «миксбордер»*** (Метро, Санкт-Петербург; 23.10.2006).

Полная представленность темпоральной цепочки — явление не столь частое в корпусе высказываний. Чаще вербализируется темпоральный уточнитель *сейчас*, а вторая единица как более привычная и знакомая вводится без уточнителя:

(18) *У Сергея Кириченко самое что ни на есть мужское увлечение —* ***пауэрлифтинг****. Этим* ***модным словом*** *сейчас называют* ***силовое троеборье*** (Московский железнодорожник; 12.10.2011).

(19) *Важны были искренность и* ***творческий кураж****, то, что называют нынче* ***модным словом «креатив»*** (Московский Комсомолец; 28.05.2012).

В анализируемом корпусе присутствуют усеченные варианты метаязыковых высказываний (без темпоральных уточнителей, а иногда и без метаоператора *модное слово*) с фокусировкой внимания на сопоставлении двух разновременны́х единиц:

(20) *Одно то, что в сборной начались* ***процессы,*** *описываемые* ***модным словом «движуха»****, уже хорошо* (Спорт-экспресс; 16.01.2012).

(21) ***Менеджер по клинингу*** (***в переводе на старорусский — уборщица***) (Секретные материалы (Санкт-Петербург)*; 05.08.2013).*

(22) ***Форсайт*** *— это, пожалуй, то, что в 90-е годы было принято называть другим* ***модным словом — «стратегическое мышление».*** *Не секрет, что мы живём в эпоху, когда* ***парикмахеры*** *стали* ***стилистами****,* ***художники — дизайнерами, уборщицы — клининг-персоналом, приказчики — менеджерами*** (Виртуальный Иркутск, г. Иркутск; 10.01.2007).

В группу усеченных метавысказываний можно включить также контексты с указанием только на одну, вышедшую из моды, лексическую единицу в составе оппозитивной пары:

(23) *Крылов и Гусев — можно сказать, еще не вполне оперившиеся* ***конферансье*** (***извините за старомодное слово***)*, хоть и с птичьими фамилиями* (Московская правда / Музыкальная правда N13; 02.04.2004).

(24) *К примеру, еще новичком на Рауэнштайнгассе Мадзини употреблял* ***старомодные слова*** *вроде* ***«синематограф»****, говорил* ***«на сей конец»****,* ***«высокие помыслы»****,* ***«следственно»*** *или* ***«телефонировать»*** (Иностранная литература; 14.02.2003).

Велика доля высказываний, фиксирующих смену номинаций, с метаоператором ***как сейчас модно говорить:***

(25) ***Оптимизировать, как модно говорить*** *сейчас, или* ***секвестрировать, как модно было говорить*** *чуть раньше* (Неделя. Подмосковье; 21.04.2010).

(26) *Государство выступает в роли идеолога и двигателя* ***тенденций*** *или,* ***как сейчас модно говорить****, восходящих* ***трендов*** (Известия Калининграда, Калининград; 26.01.2006).

С помощью корпусов при поддержке метаязыковых показателей возможна попытка определения, на какой этапе вхождения новой единицы в широкий речевой обиход носитель языка осознает смену номинаций.

Обратимся к анализу контекста (26). Первое метавысказывание о смене номинаций появилось в 2006 году, количество употреблений лексемы *тренд* в этом году составило 5,5 % от общего количества, соответственно — лексемы *тенденция* — 94,5 %. Пик модного цикла — самая высокая частотность метавысказываний *модное слово тренд* — приходится на 2011 год (при этом количество употреблений: *тренд* — 16 %, *тенденция* — 84 %).

Мы провели такой же подсчет с парой *лузер — неудачник*. Первая фиксация смены слов датируется 2005 годом:

(27) *Последнее время* ***то тут, то там мелькает слово «лузеры»****, которое понимается как* ***«неудачник»*** (Россія; 26.05.2005).

Процентное соотношение употреблений: 2005 год — *лузер* 6,5 %, *неудачник* — 93,5 %. Пик модного цикла — самая высокая частотность метавысказываний *модное слово лузер* приходится на 2011 год:

(28) *А жесткое заморское* ***слово «лузер»****, сменившее жалостливое «неудачник»?* (Версия, Башкортостан; 15.07.2011).

Процентное соотношение употреблений: 2011 год — *лузер* 16,3%, *неудачник* — 83,7%.

Опыт количественных подсчетов с учетом метавысказываний позволяет сформулировать предварительный вывод: носителем языка смена номинаций начинает осознаваться на рубеже 5–6 процентной отметки, пик модного цикла слова осознается от 16%. Безусловно, такой сопоставительный подход нуждается в дополнительном исследовании на репрезентативном материале.

2.4. Демонстративность модного слова

Последний параметр моды — демонстративность — не ограничен временны́ми и пространственными рамками. Демонстративность «имеет корни в биологических аспектах человеческого существования» [Гофман 2000: 26–27]. Для модного объекта признаки *быть* и *казаться* практически совпадают, мода всегда на виду. Обратимся к выявлению языковых особенностей, делающих лексическую единицу ярким речевым сигналом. Обычно носитель языка воспринимает слово как неразрывное единство формального и содержательного. Но это единство в языковом сознании имеет относительный характер. Безусловно, «одежкой» слова («По одежке встречают»), привлекающей внимание носителя языка, является формальная сторона знака, фоника слова. Выделим контекстные маркеры, которые фиксируют черты формальной необычности, отступления от стандартного восприятия слова:

(29) *Мы уже забыли начало 90-х годов прошлого века, когда* ***в моду вошло*** *непривычное для нашего слуха* ***словечко «кондоминиум»*** (Строительная газета, Москва; 27.07.2012)

(30) *А еще сейчас одно* ***модное слово*** *появилось, вообще хрен выговоришь, только если по слогам:* ***само-и-ден-ти-фикация!*** (Москва; 15.11.2002)

(31) *Но есть, есть чужеземные* ***модные слова*** *настолько неприятные, что так и хочется их выплюнуть. Ну, например,* ***гаджеты***. (Московский комсомолец; 30.03.2013)

(32) *На влюбленных лопухов яркие* ***модные названия цвета типа «мурена», «лагуна», «чароит»*** *действуют завораживающе* (Комсомольская правда; 22.04.2000).

Первая группа лексем, которая выделяется на основании признака формальной яркости и необычности, — это заимствованные неологизмы, семантическая неосвоенность которых усиливается «чужеродностью» формы. Работы по фоносемантике свидетельствуют, что фоника имени создается относительной краткостью слов, сочетаемостью и качеством составляющих названия звуков. В русле общей тенденции слоговой структуры русского языка в качестве идеальных

информантами оценивались короткие названия в 2–3 слога [Копачева 1990: 90]. Длинные названия даже в интервале 4–5 слогов нежелательны. Метавысказывания (29, 30) демонстрируют подобный вариант. Предпочтения отдаются ассоциативным возможностям сонорных, гласных переднего ряда и гласному «а», которые названы «легкими», «светлыми» звуками, отсюда привлекательность контекста (32). Негативное восприятие иностранного слова провоцируют нежелательные ассоциации с русскими словами отрицательной семантики (32), а также не характерные для русского языка стечения согласных, зияния гласных, одинаковые по артикуляции звуки и т. д. Отсюда в метавысказываниях по поводу модных заимствований выделяется речевой шлейф сильных и неоднозначных аксиологических реакций: от привлекательности, ореола престижности иностранного слова до ксенофобии, отрицательного отношения ко всему «чужому».

Признаком внешней привлекательности обладают и другие группы лексики, к которым можно отнести, в частности, стилистически окрашенную лексику и окказиональные слова.

Заключение

При обыденном понимании модного слова (то, что дает корпус метаязыковых высказываний в целом) класс анализируемых единиц достаточно велик. В него включается любая единица, отвечающая двум ценностным параметрам модного объекта — современности и универсальности. Требование одновременной реализации в слове всех критериев ценностей моды сужает класс искомых лексем. В него могут войти лишь высокочастотные актуальные единицы, заменяющие старые номинации и при этом обладающие яркой «упаковкой». С опорой на приведенный выше анализ вербализованных показаний метаязыкового сознания представляется, что класс модных лексем можно представить как полевую структуру с ядром слов, удовлетворяющих всем предъявляемым ценностным критериям модного объекта, и обширной периферией. К периферии относятся актуализированные высокочастотные единицы. Сочетание двух подходов к изучению модных слов — количественной обработки материала и показаний метаязыкового сознания — позволит пролить свет на ряд проблем, связанных с этапами социализации слова в языке.

Литература

1. *Барт Р.* (2003), Система моды. Статьи по семиотике культуры. М., Изд-во им. Сабашниковых.
2. *Барт Р.* (1997), Дендизм и мода, Художественный журнал, № 18, с. 6–8.
3. *Бодрийяр Ж.* (2000), Символический обмен и смерть. М., Добросвет.
4. *Бічай Ю. В.* (2003), «Модні» слова в сучасній російській мові (на матеріалі тлумачних словників і мовленисвої практики масс-медіа кінця XX — початку XXI ст.) Автореферат дисертації на здобуття наукового ступеня кандидата філологічних наук. Дніпропетовськ.

5. *Булыгина Т. В., Шмелев А. Д.* (1998), Folk linguistics // Русский язык в его функционировании. Тез. докл. международн. конференции. Третьи Шмелевские чтения. 22–24 февраля 1998 г. М., с. 13–15.
6. *Гофман А. Б.* (2000), Мода и люди. Новая теория моды и модного поведения. 2-е изд. М., Агентство «Издательский сервис», «Издательский ГНОМ и Д».
7. *Дергунова Л. А.* (2004), Мода как социальная технология. Шахты, РГУЭС.
8. *Килошенко М.* (2001), Психология моды. СПб., Речь.
9. *Копочева В. В.* (1993), Оценка названия, Детерминационный аспект функционирования значимых единиц языка: языковые и неязыковые факторы. Барнаул, Изд-во Алтайского гос. университета, с. 88–95.
10. *Мустайоки А., Вепрева И. Т.* (2006), Какое оно, модное слово: к вопросу о параметрах языковой моды, Русский язык за рубежом, № 2, с. 45–62.
11. *Михалева У.* (2010), Система моды. М., Российская политическая энциклопедия.
12. *Новиков Вл.* (2005), Словарь модных слов. М., Зебра Е.
13. *Новиков Вл.* (2012) Словарь модных слов: Языковая картина современности. М., АСТ–ПРЕСС КНИГА.
14. *Орлова Л. В.* (1989), Азбука моды. М., Просвещение.
15. *Розен Е. В.* (1991), Новые слова и устойчивые словосочетания в немецком языке. М., Просвещение.
16. *Такер Э.* (2003), История моды. М., Аст, Астрель.
17. *Титкова О. И.* (1998), Тенденции развития модных слов в лексиконе современного немецкого языка (70–90-е гг.) // Терминоведение, № 1–3, с. 150–157.
18. *Федорова Л. Л.* (ред.) (2014) Мода в языке и коммуникации. М., Изд. центр РГГУ.
19. *Ятина Л. И.* (2004), Мода глазами социолога: результаты эмпирического исследования // Социология и социальная антропология. Т. 1, № 2, с. 121–133.
20. *Hoenigswald H.* (1966), A proposal for the study of folk linguistics, Sociolinguistics: Proceedings Of the UCLA Conference, 1964, The Hague, Paris, pp. 16–21.
21. *Mjos O. J., Moe H., Sundet V. Sch.* (2014), The functions of buzzwords: A comparison of ‟Web 2.0’ and ‟telematics’. First Monday. Peer-reviewed Journal on The Internet. Vol. 19, Number 12–1, December.
22. *Tsun-Jui L., Shu-Kai Hs., Prevot L.* (2013), Observing Features of PTT Neologisms: A Corpus-driven Study with N-gram Model Proceedings of the Twenty-Fifth Conference on Computational Linguistics and Speech Processing (ROCLING 2013). — Режим доступа: http://anthology.aclweb.org/O/O13/O13-1025.pdf (дата обращения: 4.02.15).

References

1. *Bart R.* (1997), Dandyism and fashion [Dendizm i moda], Art Magazine [Hudozhestvennyj zhurnal], Vol.18, pp. 6–8.
2. *Bart R.* (2003), The system of fashion. Articles on semiotics of the culture [Sistema mody. Stat’i po semiotike kul’tury], Izdatel’stvo imeni Sabashnikovyh, Moscow.

3. *Baudrillard J.* (2000), Symbolic exchange and a death [Simvolicheskiy obmen i smert'], Dobrosvet, Moscow.
4. *Bichay J. V.* (2003), «Modern» words in the current Russian language (on the material of explanatory dictionaries and the speech patterns of mass-media of the late XXth — early XXIst centuries) [«Модні» слова в сучасній російській мові (на матеріалі тлумачних словників і мовленисвої практики масс-медіа кінця XX–початку XXI ст.)], PhD Thesis [Автореферат дисертації на здобуття наукового ступеня кандидата філологічних наук], Dnipropetrovsk National University, Dnipropetrovsk.
5. *Bulygina T. V., Shmelev A. D.* (1998), Folk linguistics, The Russian language in its functioning: Proceedings of the International Conference «The third Chmielevski's readings. February 22–24, 1998» [Russkij yazyk v ego funktsionirovanii. Tez. dokl. mezhdunarodn. konferencii. Tret'i Shmelevskie chtenija. 22–24 fevralja 1998 g.], Moscow, pp. 13–15.
6. *Dergunova L. A.* (2004), Fashion as a social technology [Moda kak sotsial'naya tehnologiya], RGUES, Shahty.
7. *F'odorova L. L.* (ed.) (2014), Fashion in the language and communication [Moda v yazyke i kommunikatsiy], Publ.center RGGU, Moscow.
8. *Gofman A. B.* (2000), Fashion and people. A new theory of fashion and fashionable behavior [Moda i lyudi. Novaja teoriya mody i modnogo povedeniya], 2nd ed., Publ. Service, Publ. GNOM i D, Moscow.
9. *Hoenigswald H.* (1966), A proposal for the study of folk linguistics, Sociolinguistics: Proceedings Of the UCLA Conference, 1964, The Hague, Paris, pp. 16–21.
10. *Kiloshenko M.* (2001), The psychology of fashion [Psikhologiya mody], Rech', St.-Petersburg.
11. *Kopocheva V. V.* (1993), Name evaluation [Otsenka nazvaniya], Determination aspect of the meaningful units of language: linguistic and non-linguistic factors [Determinatsionnyy aspect funktsionirovaniya znachimykh edinits yazyka: yazykobye i neyazykovye factory], Altaian university, Barnaul.
12. *Mikhaleva W.* (2010), The system of fashion [Sistema mody], Russian Political Encyclopedia, Moscow.
13. *Mjos O. J., Moe H., Sundet V. Sch.* (2014), The functions of buzzwords: A comparison of "Web 2.0' and "telematics', First Monday, Peer-reviewed journal on the Internet, Vol. 19, Number 12–1, December.
14. *Mustajoki A., Vepreva I. T.* (2006), What is it, a fashionable word: the question of the language fashion options [Kakoe ono, modnoe slovo: k voprosu o parametrah yazykovoy mody], The Russian language abroad [Russkiy yazyk za rubezhom], Vol. 2, pp. 45–62.
15. *Novikov Vl.* (2005), Dictionary of fashionable words [Slovar' modnykh slov], Zebra, Moscow.
16. *Novikov Vl.* (2012), Dictionary of fashionable words: A language picture of a modern world [Slovar' modnykh slov: Yazykovaya kartina sovremennosti], AST-PRESS BOOK, Moscow.
17. *Orlova L.* (1989), The ABC of fashion [Azbuka mody], Prosveshchenie, Moscow.
18. *Rosen E. V.* (1991), New words and set phrases in the German language [Novye slova i ustoychivye slovosochetaniya v nemetskom yazyke], Prosveshchenie, Moscow.

19. *Titkova O. I.* (1998), Trends in buzz words in the lexicon of the modern German language (70–90-ies.) [Tendentsii razvitija modnykh slov v leksikone sovremennogo nemetskogo yazyka (70–90-e gg.)], Terminology [Terminovedenie], Vol. 1–3, pp. 150–157.
20. *Tsun-Jui L., Shu-Kai Hs., Prevot L.* (2013), Observing Features of PTT Neologisms: A Corpus-driven Study with N-gram Model Proceedings of the Twenty-Fifth Conference on Computational Linguistics and Speech Processing (ROCLING 2013), available at: www // anthology. aclweb. org/O/O13/O13-1025.pdf.
21. *Tucker E.* (2003), The history of fashion [Istoriya mody], Ast, Astrel, Moscow.
22. *Yatina L. I.* (2004) Fashion by the eyes of sociologist: results of the empirical research [Moda glazami sotsiologa: rezul'taty èmpiricheskogo issledovaniya], Sociology and Social Anthropology [Sotsiologiya i social'naya antropologiya], Vol. 1(2), pp. 121–133.

ГРАФОВЫЙ ПОДХОД В ЗАДАЧЕ ПОСТРОЕНИЯ СИНТАКСИЧЕСКИХ ДЕРЕВЬЕВ ДЛЯ РУССКОГО ЯЗЫКА

Музычка С. (s.muzychka@samsung.com),
Пионтковская И. (p.irina@samsung.com)

Исследовательский Центр Самсунг, Москва, Россия

Синтаксический анализ является одним из ключевых компонентов в большом количестве задач автоматической обработки текстов на естественном языке. Решение задачи построения деревьев зависимостей необходимо достаточно широкому кругу систем машинного перевода, автоматического синтеза и распознавания речи и пр. В статье изложен метод автоматического построения синтаксических деревьев на основе графового (graph-based) подхода. Мы предлагаем достаточно простую для реализации вероятностную модель, являющуюся модификацией работы [3]. Наш подход состоит в замене статистической суммы (partition function) некоторым приближением, не ухудшающим качество алгоритма и существенно снижающим сложность вычислений. Продемонстрировано, что указанный метод может быть успешно применен к синтаксическому анализу национального корпуса русского языка SynTagRus. Полученная точность алгоритма (по UAS) превосходит существующие аналоги.

Ключевые слова: синтаксический анализ, графовый подход, SynTagRus, графические модели

GRAPH-BASED APPROACH IN THE DEPENDENCY PARSING TASK FOR RUSSIAN LANGUAGE

Muzychka S. (s.muzychka@samsung.com),
Piontkovskaya I. (p.irina@samsung.com)

Samsung R&D Institute, Moscow, Russia

Dependency parsing is one of the key components in a large number of tasks of automatic processing of natural language texts. Effective dependency tree construction can be applied to a wide variety of machine translation systems, automatic speech synthesis and recognition, and so forth. Graph-based approach in dependency parsing proved to be efficient for morphologically rich languages due to its possibility to deal with non-projective dependency trees and flexible word order. Usually graph-based methods

enable to perform probabilistic analysis over distribution on the set of syntax trees. In some NLP tasks it is not required to present a full syntactic parsing (in particular, to set labels on the edges of the tree). It is enough to find a parent for a given token. In this case, the graph-based approach is more appropriate because the likelihood that a token is an ancestor of the other, can be calculated by the explicit formula.

We consider a task of automatic syntax tree construction with application to Russian language corpus SynTagRus. We propose a novel technique which enables to reduce time costs for training and doesn't affect resulting accuracy. Experiments show that our algorithm outperforms existing analogues on SynTagRus in UAS (unlabeled attachment score) measure (percentage of correctly identified unmarked dependencies).

Key words: syntax parsing, graph-based approach, SynTagRus, dependency parsing

1. Introduction

Dependency parsing is one of the key components in a large number of tasks of automatic processing of natural language texts. An automatic construction of syntactic trees can be applied to a wide variety systems of machine translation, automatic speech synthesis and recognition.

Existing methods of dependency parsing can be divided into two fairly large groups: transition-based and graph-based approaches [4]. The first is based on the construction of a finite state automaton. Tokens of the sentence are sequentially fed to the algorithm after which the automaton transfers to a new state and corrects syntax tree. One popular example of this approach is MaltParser, which has been successfully applied to the analysis of many languages with complex morphology, in particular for the Russian language [7, 8].

Graph-based approach is based on a discriminative probabilistic model of dependencies between tokens of the sentence. Its popular implementation is MSTParser [5]. The advantage of graph-based approach towards a transition-based parsing is the capability of non-projective trees construction without any processing.

Usually graph-based methods enable to perform probabilistic analysis over distribution on the set of syntax trees. In some NLP tasks it is not required to present a full syntactic parsing (in particular, to set labels on the edges of the tree). It is enough to find a parent for a given token. In this case, the graph-based approach is more appropriate because the likelihood that a token is an ancestor of the other, can be calculated by the explicit formula [3].

We propose a simple for implementation probability model on the basis of graph-based approach which is a modification of [3]. The approach is based on the replacement of the partition function with its approximation which doesn't affect the performance of the algorithm and reduces computation costs. Out algorithm can be successfully applied to Russian language corpus SynTagRus. The accuracy of the resulting classifier outperforms existing analogues in UAS (unlabeled attachment score) measure [7, 8].

2. Model Description

Dependency tree T of the sentence $x = \{x_i\}_{i=1}^{N}$ consisting of N tokens x_i is an oriented graph-tree which nodes correspond to x and one outstanding node without incoming edges is called root $r(T)$. Denote the set of all dependency trees of sentence x by $\mathcal{T}(x)$. The problem is to construct the most appropriate syntax tree $T \in \mathcal{T}(x)$ for a given sentence x. Everywhere below for simplicity we identify each token with its index in the sentence i.

The following probability model was suggested by [3]. For each pair of tokens (h, m), $h \neq m$ (h—head, m—modifier) we construct a set of features $\mathbf{f}_{hm}$ (for example, $\mathbf{f}_{hm}$ can contain parts of speech of h and m, distance between them or lemmas of their neighbours) and calculate its pairwise potential

$$w_{hm} = exp\left(\sum_{f \in \mathbf{f}_{hm}} \theta_f\right) \text{ ($\theta_f \in \mathbb{R}$ are parameters of the model)}$$

measuring the rate of dependence between h and m. Also for each token m we construct a set of features $\mathbf{f}_m$ and calculate single potential

$$w_{0m} = exp\left(\sum_{f \in \mathbf{f}_m} \theta_f\right)$$

(index 0 expresses fictitious node-root). The weight $w(t)$ of dependency parsing tree T is defined by

$$w(T) = w_{0r(T)} \prod_{(h,m) \in E(T)} w_{hm},$$

where $r(T)$ is the tree root, and $E(T)$ is the set of all edges of T. The probability of the tree $P(T|x)$ is a normalized weight

$$P(T|x) = \frac{w(T)}{Z_0}, \text{where } Z_0 = \sum_{T \in \mathcal{T}(x)} w(T) \quad \text{is its partition function} \tag{1}$$

Having trained model the most appropriate dependecy tree can be found using Eisner [2] or Chu-Liu-Edmonds algorithms [1].

The parameters of the model $\{\theta_f\}$ are fitted on the training corpus using maximum likelihood method. Applying matrix tree theorem [9] we get that Z_0 is equal to the determinant of the matrix [3]

$$\begin{pmatrix} w_{01} & w_{02} & w_{03} & & w_{0N} \\ -w_{21} & \sum_i w_{i2} & -w_{23} & & -w_{2N} \\ -w_{31} & -w_{32} & \sum_i w_{i3} & & -w_{3N} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ -w_{N1} & -w_{N2} & -w_{N3} & & \sum_i w_{iN} \end{pmatrix}$$

Therefore the most computationally difficult problem is to calculate logarithm of partition function (1) and its gradient. To deal with this obstacle we use 2 technical tricks (see appendix for proof)

1. Replace Z_0 with the sum over all oriented graphs

$$Z_1 = \prod_{\substack{(h.m):h\neq m \\ 0\leq h\leq N \\ 0<m\leq N}} (1 + w_{hm}) \tag{2}$$

2. Replace Z_0 with the sum over all functional graphs (oriented graph is functional if number of input edges equals 1 for each node in the graph

$$Z_2 = \prod_{m=1}^{N} \sum_{\substack{(h.m):h\neq m \\ 0\leq h\leq N}} w_{hm} \tag{3}$$

Application of (2) and (3) sufficiently reduces computational complexity during training and doesn't decrease accuracy of the resulting algorithm as experiments show.

3. Results

The experiments were carried out on the SynTagRus corpus [7]. We used different combinations of parts of speech, words and lemmas as features. Let *pos*(*h*) be a part of speech of token h; also denoted by *morph*(*h*) its full morphological label; *lemma*(*h*)—its lemma; *token*(*h*)—its token and *dist*(*h,m*)—distance between tokens *h* and *m*. In the following table we present the list of the most significant features used in our model

Features for w hm	Features for w hm
pos(*h*) + *pos*(*m*)	*pos*(*h*)
morph(*h*) + *morph*(*m*)	*token*(*h*)
token(*h*) + *pos*(*m*)	*lemma*(*h*)
pos(*h*) + *token*(*m*)	*morph*(*h*)
lemma(*h*) + *pos*(*m*)	
pos(*h*) + *lemma*(*m*)	
"*head=*" + *pos*(*h*)	
"*modifier=*" + *pos*(*m*)	
dist(*h, m*)+*pos*(*h*)+*pos*(*m*)	
dist(*h,m*)	

Also we used features taking into account punctuation, frequent tokens and neighbours of h and m.

For training and testing the corpus was separated into 2 parts by random split: 90%—for training and 10%—for testing. We consider 3 variants of training corresponding to the choice of partition function and 2 variants of inference: Eisner and Chu-Liu-Edmonds algorithms. The results of the testing (by UAS measure) are presented in the following table.

	Z_0	Z_1	Z_2
Eisner (perfect morphology)	90.50%	90.29%	**90.56%**
Chu-Liu-Edmonds(perfect morphology)	90.09%	90.03%	90.31%
Eisner (predicted morphology)	86.76%	86.54%	86.97%
Chu-Liu-Edmonds (predicted morphology)	86.05%	85.92%	86.46%

Morphological tagging was performed with morphological parser [6]. The following table shows the results of error statistics depending on part of speech of dependent word for the best experiment (perfect morphology + Eisner + Z_2)

Part of speech	Accuracy on POS	Common mistake
S	93.80%	2.490%
PR	82.57%	2.020%
V	88.90%	1.600%
CONJ	80.67%	1.240%
ADV	85.87%	0.850%
A	95.83%	0.600%
PART	90.30%	0.400%
NID	87.39%	0.050%
NUM	94.48%	0.110%
UNKNOWN	76.41%	0.050%
COM	85.71%	0.002%

In order to compare our accuracy with the current state-of-the-art we present the following table based on the papers [7, 8].

Nivre	Sharoff
89.00%	89.40%

4. Discussion

The accuracy of resulting classifier outperforms existing analogues [7, 8] by UAS-measure. The choice of partition function doesn't affect an accuracy of the algorithm. Moreover usage of Z_1 and Z_2 sufficiently increase efficiency of the algorithm.

References

[1] *J. Edmonds,* (1967), Optimum branching, Journal of Research of the National Bureau of Standards, 71B, pp. 233–240.

[2] *J. Eisner,* (1996), Three new probabilistic models for dependency parsing: An exploration, Proceedings of the 16th conference on Computational linguistics, Copengagen, pp. 340–345.

[3] *T. Koo, A. Globerson, X. Careras, M. Collins,* (2007), Structured Prediction Models via the Matrix Tree Theorem, Proceedings of EMNLP, Prague, pp. 141–150.

[4] *S. Kubler, R. McDonald, J. Nivre,* (2009), Dependency parsing, Morgan & Claypool publishers.

[5] *R. McDonald, F. Pereira, K. Ribarov, J. Hajic,* (2005), Non-projective Dependency Parsing using Spanning Tree Algorithms, Proceedings of HLT/EMNLP, Vancouver, pp. 523–530.

[6] *S. Muzychka, A. Romanenko, I. Piontkovaskaya,* (2014), Conditional Random Field for morphological disambiguation in Russian, Proceedings of the International conference "Dialogue 2014" Bekasovo, 2014, pp. 456–465.

[7] *J. Nivre, I. Boguslavsky, L. Iomdin,* (2008), Parsing the SynTagRus Treebank of Russian, Proceedings of the International conference "Dialogue 2008", Bekasovo, pp. 641–648.

[8] *S. Sharoff,* (2011), The proper place of men and machines in language technology: Processing Russian without any linguistic knowledge, Proceedings of the International conference "Dialogue 2011", Bekasovo.

[9] *W. Tutte,* (1984) Graph Theory, Addison-Wesley, Menlo Park, 1984.

Appendix

Proof of (2) Denote by the set of all pairs $(i,j), i \neq j, 0 < i \leq N, 0 < j \leq N$. Opening the brackets we get

$$Z_1 = \prod_{\substack{(h.m):h\neq m \\ 0\leq h\leq N \\ 0<m\leq N}} (1 + w_{hm}) = \sum_{V\subset U} \prod_{(h,m)\in V} w_{hm}$$

Each term in the right part corresponds to the weight of the oriented graph with edges $(h,m) \in V$. Since each corresponding graph is present in the sum only once we get the result.

Proof of (3) Opening the brackets we get

$$Z_2 = \prod_{m=1}^{N} \sum_{\substack{(h.m):h\neq m \\ 0\leq h\leq N}} w_{hm} = \sum_{h_1\neq 1} \sum_{h_2\neq 2} \dots \sum_{h_N\neq N} (w_{h_1 1}\, w_{h_2 2} \dots w_{h_N N})$$

Each term in the right part corresponds to the weight of the corresponding functional graph. Since each corresponding graph is present in the sum only once we get the result.

COREFERENCE CHAINS IN CZECH, ENGLISH AND RUSSIAN: PRELIMINARY FINDINGS[1]

Anna Nedoluzhko (nedoluzko@ufal.mff.cuni.cz)[1],
Svetlana Toldova (stoldova@hse.ru)[2],
Michal Novák (mnovak@ufal.mff.cuni.cz)[1]

[1]Charles University in Prague
[2]Russian Research University "Higher School of Economics"

This paper is a pilot comparative study on coreference chaining in three languages, namely, Czech, English and Russian. We have analyzed 16 parallel English-Czech newspaper texts and 16 texts in Russian (similar to the English-Czech ones in length and topics). Our motivation was to find out what the linguistic structure of coreference chains in different languages is and what types of distinctions we should take into account for advancing the development of systems for coreference resolution. Taking into account theoretical approaches to the phenomenon of coreference we based our research on the following assumption: the recognition of coreference links for different structural types of noun phrases is regulated by different language mechanisms. The other starting point was that different languages allow pronominal chaining of different length and that coreference chains properties differ for the languages with different strategies for zero anaphora and different systems for definiteness marking. This work reports our first findings within the task of the structural NP types' distribution comparison in three languages under analysis.

Keywords: coreference, coreference resolution, zero anaphora, NP-structural types distribution, cross-language comparison

> "Statistical models of anaphora resolution so far have only scratched the surface of the phenomenon, and the contributions to the linguistic understanding of the phenomenon have been few". (Poesio et al., 2011)

1. Introduction

Coreferential—anaphoric in particular—relations are common to most languages. However, the means of expressing these relations in languages can be different. The use of anaphoric expressions is influenced by different factors. These factors are not limited to such relatively language-independent ones, as context, pragmatic situation and semantics of referring expressions, but also by language-dependent

[1] The reported study was partially supported by RFBR, research project No. 15-07-09306

factors, such as pro-drop character of a language, different kinds of syntactic constructions common for a language in question, and so on. The present research is a pilot study aiming at comparison of the coreference chains structure in three languages—English, Czech and Russian. The goal of our research is twofold. From the linguistic point of view, hypotheses on coreferential expressions and coreferential chains will be formulated. From a computational perspective, this work will help us find specific features related to anaphoric expressions in text that can be further used as background knowledge for the development of a multilingual tool for coreference and anaphora resolution. In this paper, we will try to find out whether there are any language specific parameters of coreference chaining in different languages, what they are in particular, and what language phenomena they could be accounted for.

By a corefential chain we mean all the mentions of one and the same entity (or in some case mental entity or notion) through the whole text, irrespective of the features of a particular noun phrase (NP):

(1) (English) *Police say a husband fatally shot his wife and another man before* [$\varnothing_{PRO}$] *killing himself in a central Pennsylvania motel room.* […] *The York County Coroner's Office says 35-year-old Donnell Graham shot his wife.*

(2) (Russian) … *banker Lamberto Albuccani zastrelil svoju zhenu,…, a potom* [$\varnothing_{PRO}$] *zastrelilsya sam.*
[*lit.* The banker Lamberto Ambuccani shot his wife and then shot himself]

In (1), all the underlined NPs refer to the same entity (Donnell Graham), the first NP is an indefinite description denoting an entity's social role (*husband*), then we have a possessive anaphoric pronoun *his*, a reflexive pronoun *himself*, a zero inexpressible pronoun $\varnothing_{PRO}$[2] as a subject of a clause with a gerund *killing*, and a full NP including a proper name in the second sentence (*35-year-old Donnell Graham*). As for Russian Example (2), we have the proper name within a full NP (*banker Lamberto Albuccani*), a possessive reflexive svoju 'his', a zero pronoun as an agent of a finite verb *zastrelilsya*.

2. Motivation

Two basic issues, determining different parameters of coreferential expressions and chains, served us as motivation for the beginning of our research.

The first issue concerns the structural types of coreferring expressions. Although the quality of coreference resolution is usually evaluated as a whole, some researchers claim that syntactic and semantic structure of coreferring expressions might influence crucially the quality of coreference relations extraction. [Poesio et al. (2011)] suggests that recognition of different types of referring expressions (e.g., syntactic

[2] For zero NPs and their status see Section 2.

anaphora, named entities, metaphors, etc.) should be evaluated separately. We consider this approach very reasonable. Indeed, we should take into account that the referential choice for different NP classes (e.g., reflexive pronouns vs. anaphoric ones) is regulated by different language mechanisms that are studied within quite different linguistic paradigms. For example, syntactic anaphora (e.g., bound anaphora such as reflexives, some types of zero pronouns) is in the domain of formal syntactic investigation (see, e.g., in Chomsky (1981), Reuland (2011)). The discourse anaphora (i.e. 3rd person pronouns or demonstratives) is in the focus of referent activation theories (see, for example, Givon (1983), Ariel (2001), Kibrik (1997)). The recognition of metaphoric and metonymic expressions within coreference chains includes the semantic similarity detection (e.g., NP 'treasure' used to refer to a golden ring). To sum up, there are different language mechanisms, responsible for different NP types referential choice. Thus, the data on different NP types distribution within coreference chains for a particular language might be useful both for coreference chaining theoretical issues and for the multilingual coreference resolution task.

The second issue concerns language-specific zero pronominal elements in particular. For example, in our case, the three languages under discussion differ in the Subject ellipsis[3] options. In English, the syntactic subject should be always expressed explicitly. In other languages, like Czech and Russian, the subject can be omitted, or, in other words, free zero-pronoun (∅)[4] is used in some contexts (see Examples (3a-c)):

(3) *a.* (English) *Peter came home. *(∅) Watched TV and went to sleep.*
b. (Czech) *Petr se vrátil domů. ∅ Podíval se televizi a šel spát.*
c. (Russian) *Petya prishel domoj.* (On/∅) *Posmotrel televisor i poshel spat.*

We should also take into account another case of non-overt Subjects, coreferential to a NP in a previous context, that is the distribution of unexpressed Null Subjects (PRO) regulated by syntactic rules (e.g., PRO in infinitival constructions). To sum up, the syntactic properties of a language can influence the coreference chaining distribution in this language.

Thus, the purpose of our pilot study is just to compare the distribution of NP structural types (including free zero pronouns) in coreference chains for three languages.

3. Theoretical Background

In theoretical linguistics, the analysis of coreferential chains most closely relates to referent activation theories (see, e.g., Givon, 1983; Ariel, 2001; Kibrik, 2011; Kibrik, 1997, etc.). These studies suggest the model of referential choice (the choice of a particular NP type) based on the degree of referent salience. They mostly address this phenomenon in one language. Some studies analyze the predictability

[3] Languages like Czech are called pro-drop languages. In pro-drop languages a pronoun (primarily in subject position) could be omitted in contexts where they could be pragmatically inferred.

[4] We do not draw a distinction between zero-pronoun and ellipsis here

of upcoming referents in relation to the choice of coreferring expressions and its status in information structure of an utterance (see the algorithm, determining the degree of salience in Hajičová et al., 2006; Lambrecht, 1994; Strube—Hahn, 1999, etc.). The referential choice regulation in subject position for Russian and its comparison to other world languages, first of all, to Germanic ones, is provided in Kibrik (2013). The deeper diachronic analyses of subject reference in Russian can be found in Sidorova (2013) and Kibrik (2013). A contrastive research of coreference and anaphoric reference of demonstratives in French and Portuguese is presented in Salmon-Alt et al. (2005).

As for corpus approaches, there is a large amount of large-scale annotated data for coreference, anaphoric relations, event anaphora (or discourse deixis, reference to events), bridging relations (associative anaphora) and so on. However, as far as we know, there is a very little number of studies, analyzing the difference between anaphoric expressions, based on large-scale annotated parallel corpora. The comparison of pronominal and zero coreferential expressions in Czech and English has been recently provided in Novák—Nedoluzhko (forthcoming in 2015). However, this work focuses on mappings between certain classes of coreferential expressions, and it does not take into account the structure of coreferential chains as a whole. Conversely, coreference chains have been included in the statistical analysis of cohesive devices in Kunz et al. (2015) for German-English corpus, containing written and spoken texts (GECCo), where the number of chains and chain lengths have been computed, but the collected numbers have not been analyzed yet.

4. Data

4.1. Description of the Corpora

Prague Czech-English Treebank (PCEDT) is a manually parsed Czech-English parallel corpus of 1.2 million words in almost 50,000 sentences for each language. The English part consists of the Wall Street Journal (WSJ) section of the Penn Treebank (Linguistic Data Consortium, 1999). The Czech part was translated from the English source sentence by sentence. PCEDT 2.0 is annotated on three layers; the most abstract (tectogrammatical) layer includes the annotation of coreferential links. For the detailed overview of the underlying linguistic theory, see Hajič et al. (2012).

For the Russian part of our investigation, we took the data from the Russian Coreference Corpus (RuCor). The corpus is the Gold Standard corpus for coreference resolution evaluation for Russian (Toldova et al., 2014). It consists of two parts, both manually annotated for coreference: the learning set and the evaluation set, 185 texts (200 000 tokens) in total. The corpus contains automatic morphological annotation. The set of tools, developed by S. Sharoff for Russian, was used, which includes a tokenizer, a TreeTagger-based (Schmid, 1994) part-of-speech tagger, and a lemmatizer, based on CSTLemma (Jongejan—Dalianis, 2009). Some of the tagger mistakes, e.g. in anaphoric pronouns POS detection, were corrected manually.

4.2. Coreference annotation in PCEDT and RuCor

In PCEDT, coreference links are annotated (mostly manually) for both the Czech and English parts separately. The coreference annotation captures the so-called grammatical and textual coreference. The grammatical coreference typically occurs within a single sentence, with the antecedent being able to be derived on the basis of grammar rules of a given language. It includes relative pronouns, verbs of control, reflexive pronouns, reciprocity and verbal complements. On the other hand, the textual coreference is not expressed by grammatical means alone, but also via context. Annotation of textual coreference in PCEDT captures coreferential relations between personal and possessive pronouns, anaphoric zeros, noun phrases with nominal head, numerals and pronominal adverbs with demonstrative meaning (there, then, etc.). Also the cases of event anaphora (anaphoric reference to clauses) have been annotated. A detailed description of the types of grammatical and textual coreference annotated in PCEDT can be found in Nedoluzhko et al. (2014).

In RuCor, the grammatical and textual coreference was annotated manually. The following coreferential relations have been taken into consideration: reflexive and relative pronouns (included in the so-called syntactic anaphora), 3rd person pronouns and 3rd person possessive pronouns, anaphoric zero, and noun phrases of different types. There is no syntactic zero (PRO) or event anaphora annotation. More details on coreference annotation in RuCor can be found in Toldova et al. (2014).

4.3. Dataset used for the experiment

For the aim of our analysis, we have chosen comparable texts for each corpus. Taking into account the length of the texts and their genre specification, we have extracted 16 parallel English-Czech texts from PCEDT and 16 Russian texts from RuCor. The thematics, genre, type and size of texts were determinants of the excerption. The average length of the texts is 31.8 sentences for PCEDT and 31.4 for Russian, the shortest text consisting of 14 sentences and the longest one—of 64 sentences. As for genres, these are all journalistic texts; the topics are distributed as shown in Table 1[5].

Table 1. The thematic structure of texts in RuCor and PCEDT (in sentences)

	PCEDT	**RuCor**
economics texts	161	166
political news	230	231
other news	112	105
TOTAL	503	499

[5] The choice of texts has been based on the annotated data we have at our disposal. We are fully aware that the fact that Czech texts are translated from English and the Russian texts are comparable but not parallel may considerably influence the numbers.

5. Quantitative characteristics of coreference chains

We started by gathering general statistics on the number and length of coreferential chains[6]. Although these data were just the preliminary stage of our study the results have turned out to be worth of a special analysis. They are presented in Table 2.

Table 2. Number and length of coreference chains in the analyzed texts

Chain length	English	Czech	Russian
number of 2-elements chains	254 (60.5%)	304 (62.7%)	139 (52.9%)
number of chains of length 3–4	108 (25.7%)	109 (22.5%)	64 (24.3%)
number of chains of length 5–8	39 (9.3%)	47 (9.7%)	33 (12.5%)
number of chains longer than 8	19 (4.5%)	25 (5.1%)	28 (10.6%)
TOTAL number of chains[7]	**420 (100.0%)**	**485 (100.0%)**	**263 (100.0%)**

As we can see from the table, the distribution of chains with different length for Czech, English and Russian is quite similar, at least comparable. More than 50% of chains consist of two elements in all three languages, about 25% of chains consist of three or four elements, and the rest is distributed between longer coreference chains of 5 and more elements. There is a slightly stronger tendency for chains longer than 5 items in Russian. The long chains are typical for the referents that are main topics of a text. Thus, the difference could be due to the texts thematic structure (there are two texts in Russian that have 4 and 6 chains longer than 8 items).

It is noteworthy that the proportion of two-member chains in Russian is 10 percentage points less than in English and Czech. Besides, there is a substantial difference in the absolute numbers of chains. The total number of coreference chains in English and Czech is almost twice as much as the number of coreference chains in the Russian texts.

Ex facte, this difference may be due to the fact that English and Czech texts come from the parallel dataset. More than that, an attempt was made to preserve the original structure of the English texts in their Czech translations. The same approach to coreference annotation has been applied to the texts in both languages. However, even the number of chains for Czech and English differ in 15% though these texts are from the parallel datasets. Thus, the difference in chains quantity for the languages under discussion could not be accounted for by the difference in themes, referents mentioned etc. It needs some other explanation.

The preliminary analysis of sentence structure and zero pronouns distribution in three languages has drawn us to the following hypothesizes: the observed disproportion in chain numbers might be attributed to (a) more extensive use of non-finite constructions

---

[6] For all analyzed languages, coreference chain is considered to consist of all coreferential expressions including pronouns and full NPs in the gives discourse.

[7] Coreference chains statistics presented in Table 3 for English and Czech texts takes into account only those types of coreferential relations that are annotated also for Russian. For this reason, for example, syntactic (grammatical) coreference of arguments in control constructions and event anaphora (reference to sentences and clauses) have been excluded.

in English introducing elided nodes which are involved in coreference annotation (see Section 6), and even more extensive use of non-finite constructions with an inexpressible PRO in Russian, (b) and to more extensive use of free zero pronoun in Czech.

6. Qualitative characteristics of coreference chains

As claimed in typological literature (e.g. Givon, 1983), languages differ in their repertoire of the anaphoric and coreference maintenance devices, especially, pronominal and syntactic devices. For the pro-drop languages, the zero pronoun is one of the basic means of reference maintenance, while for the other languages, zero is only possible in a limited number of syntactic positions (e.g., with infinitives). Another possible difference is the frequency of anaphoric pronouns in a chain. One more point of variation is the difference in the distribution of pronominal and nominal NPs across languages. Besides, the structure of nominal NPs could vary depending on whether a language has obligatory definiteness marking, and depending on what means a language uses to compensate for the absence of grammatical definiteness.

Taking in consideration all these possibilities, we compared the structure of markables in the three languages under analysis. Table 3 shows the distribution of structural types of noun phrases in English, Czech and Russian. In the table, all markables are considered, including initial antecedents and all anaphoric mentions.

Table 3. The distribution of NP-structural types for English, Czech and Russian

NP type \ language		English		Czech		Russian	
central pronouns	Subj	121	8.6%	2	0.1%	39	3.8%
	non-subj	129	9.2%	125	7.8%	95	9.3%
Relative		96	6.9%	136	8.5%	42	4.1%
anaphoric zero		28	2.0%	304	19.1%	13	1.3%
bare noun		75	5.4%	208	13.1%	164	16.0%
NP with determiner		315	22.5%	63	4.0%	20	1.9%
NP with other modif		119	8.5%	313	19.7%	172	16.8%
NP including NE		104	7.4%	141	8.9%	80	7.8%
NE		331	23.7%	216	13.6%	379	37.0%
other		81	5.8%	83	5.2%	21	2.0%
TOTAL		1,399	100%	1,591	100%	1,025	100%

6.1. Table description and observations

Central explicit pronouns (central pronouns) include 3rd person pronouns and 3rd person possessive and reflexive pronouns that are explicitly expressed in the sentence. For subject position this group is very scarce in Czech (only two instances), thus

corroborating the pro-drop character of Czech. In Russian texts, the subject anaphora is not as rare as in Czech. However, it is substantially more frequent in English. Only 30% of Russian anaphoric central pronouns have been used in the subject position, while in English subject and non-subject anaphoric pronouns are distributed nearly equally.

Relative pronouns (relative) as anaphoric expressions in coreference chains. The number of relative pronouns in Czech coreference chains is larger than that of relative pronouns in English and in Russian. One of the reasons for this discrepancy is that English non-finite clauses are often translated to Czech as relative clauses (see Example 4)[8].

(4) *a.* (English)*: Mr. Bush had been holding out for a bill ∅ boosting the wage floor to $ 4.25 an hour.*
b. (Czech)*: Bush trval na zákoně, který by zvyšoval dolní hranici mzdy na 4.25 dolaru za hodinu.*

As will be shown in Section 7, Russian is closer to English in this respect, both of the languages preferring non-finite clauses. There are only 40 cases of clauses with the relative pronoun "kotoryj" in our texts, while there are nearly 100 participial clauses.

Anaphoric zeros (anaphoric zero) are most frequent in Czech, compensating for the lack of explicitly expressed anaphoric pronouns in the subject position, and again supporting the idea of the pro-drop character of Czech. In this respect, the results for Russian are especially interesting. Though Russian is also considered to be a pro-drop language and zero anaphora is possible in Russian (see Example 1c and 3c), there are very few examples of textual anaphoric ellipsis in Russian, not more than one or two per text. Moreover, we have no zero in the subject position? of the main clause in our text collection.

(5) *a.* (English)*: The recent explosion of country funds mirrors the "closed-end fund mania" of the 1920s* [...] *They fell into oblivion after the 1929 crash.*
b. (Czech)*: Současná exploze národních fondů je stejná jako "mánie uzavřených fondů" ve 20. letech* [...] *Po krachu v roce 1929 ∅ upadly v zapomnění.*
c. (Russian)*: Strany Nato prevoshodili Jugoslaviju …, odnako cherez 2.5 mesyaca voiny ∅ byli na predele vozmoznostej* (='NATO countries overpowered Yugoslavia …, however, after 2.5 months (they) *stretched too thin')*

The number of anaphoric zeros in English is positive even though English is not a pro-drop language. Moreover, this number seems to be relatively high. However, all 28 English anaphoric zeros in our texts are arguments of nonfinite clauses, where the syntactic subject cannot be expressed explicitly (PRO). Thus, the reason for such a high number of zeros in English is rather technical, reflecting a slightly different conception of PCEDT in understanding the distinction between PRO and ∅.

The group of **nouns** is the set of non-pronominal non-zero markables. For the analysis of coreference chains and anaphoric expressions, the following subsets

[8] On a larger corpus and with statistically more representative results, this fact was addressed in Novák—Nedoluzhko (2015).

of nominal expressions seem to be relevant: bare nouns, NPs with a determiner, NPs with other modifiers, named entities (NEs), and noun phrases that include a named entity as a dependent element (marked as NP including NE in Table 3).

Not surprisingly, the number of **bare nouns** in Russian and Czech is much larger than in English (75,208 and 164 in English, Czech and Russian, respectively). As a language with the grammatical category of definiteness, English does not use bare nouns very often. The group of coreferential English bare nouns in our text selection consists mostly of plural nouns, nouns of time (Tuesday, yesterday, etc.) that could be also considered as named entities. On the contrary, NPs with determiners prevail in English, because many elements of coreference chains in English are used with the definite article. For Czech and Russian, noun phrases with demonstratives, corresponding to "this" and "that" have been counted. The structure of NPs with other modifications need deeper investigation. These NPs can include evaluative adjectives which do not contribute to the NP definite/indefinite interpretation, or geographic and other adjectives that serve for the referent identification.

The number of **named entity** roots (**NEs**) for Czech is substantially less than in English and Russian. Even though English and Russian behave similarly in many aspects, the reason for the discrepancy between these two languages and Czech probably lies somewhere else. The high frequency of named entities in Russian may stem from the text specificity and the difference in annotation scheme. Another possible reason is the tendency in Russian to repeat full named entities in cases where Czech uses anaphoric devices. On the other hand, the difference between English and Czech is affected by the differences in sets of categories used in automatic annotation of NEs. This was provided by the tools NameTag (Straková et al., 2014) and Stanford NER (Finkel et al., 2005) for Czech and English, respectively.

The category **other** for Czech and English includes mostly coordinative and appositive structures, such as coreferential expressions and clauses (sentences, verbal phrases) as antecedents. These are also the years and **other** numerals in substantive function, local and temporal adverbs like there, then and so on.

7. Discussion

7.1. Pro-drop properties

As it has been mentioned above, one of the important differences among the three languages is the difference in zero NPs distribution. First of all, all three languages differ a lot in their pro-drop properties (see Example 1). This property is crucial for Czech: the Table 3 shows that 19% of anaphoric NPs in Czech are zeros. This affects the difference in explicit pronoun distribution for Czech and Russian. Though Russian is also a pro-drop language, the proportion of zeros is very little. Moreover, there is no ∅ in the main clause in subject position in Russian in our data. Thus, we can assume that Russian is a pro-drop language to a lesser extent than Czech. There are very few cases of zero anaphora even in subordinate clauses in Russian. Our hypothesis is that the difference in clausal structure for these two languages could play the role.

7.2. Zeros and clause structure

Another important dissimilarity in chain distribution is that the number of chains in Russian differs from those in English and Czech. In our calculations, we do not take into account the inexpressible zeros (PRO). This type of anaphora was absent in annotation scheme for RusCor. However, we can try to compensate for the lack of information on PRO distribution by taking into account the distribution of finite/non-finite forms in the languages. As far as syntactic anaphora is concerned, the distribution of syntactically regulating pronouns (PRO and some others) depends on the sentence complexity. The non-finite subordinate clauses in Russian, such as infinitival constructions or participial constructions, presuppose an inexpressible PRO in the subject position (for the PRO distribution in Russian non-finite constructions see Testelets 2001).

Thus, as it is mentioned in 5 one of our hypothesises is that the difference in coreference chaining is strongly influenced by the clause structure of a sentence.

To check this hypothesis, we have counted the number of finite and nonfinite clauses in the three languages. The total number of sentences in all three collections was approximately the same (see Table 1). The results are given in Table 4.

Table 4. Finite and nonfinite clauses in texts

	Czech	English	Russian
number of finite clauses	1,166	1,005	663
number of nonfinite clauses	97	200	379

As seen in the table, there are 379 non-finite verb forms in Russian texts (among which there are 106 participial clauses, 59 short form participles, 17 converbs and 197 infinitive clauses), thus, we expect approximately 350 PROs.

It is interesting to observe that the Czech sentence in Example 6b can be hardly reformulated using an infinite clause (it is possible, but it will be stylistically marked, see Example 6c), while in Russian, either finite subordinate clause with relative pronoun (Example 6d), infinitive (Example 6e), or infinite participial clause (Example 6f) can be used. This fact supports the hypothesis that the relatively small number of coreference chains in Russian is caused by the frequent use of nonfinite clauses in this language, the arguments of which are not annotated for coreference in Russian coreference corpus.

(6) *a.* (English) *He left a message PRO accusing Mr. Darman of selling out.*
b. (Czech) ***∅*** *Zanechal mu zprávu, ve které* ***∅*** *viní Darmana ze zaprodanosti.*
?c. (Czech) ***∅*** *Zanechal mu zprávu,* **PRO** *obviníňující Darmana ze zaprodanosti.*
d. (Russian) ***On*** *ostavil soobschenije, v* ***kotorom*** *obvinjajet Darmana v prodazhnosti.*
e. (Russian) ***On*** *ostavil soobschenije, chtoby* **PRO** *obvinit' Darmana v prodazhnosti.*
f. (Russian) ***On*** *ostavil soobschenije,* **PRO** *obvinjajuscheje Darmana v prodazhnosti.*

Some differences in the properties of coreferential chains are also caused by the differences in annotation styles, which should not be neglected. We have not addressed these in a sufficient detail, which should be one of the aims of future research.

8. Conclusions

Our pilot study has shown that the inter-language comparison of coreference chains distribution reveals a systematic difference in coreference for languages with different syntactic properties. One of very important syntactic features that should be taken into account in modelling multi-lingual anaphora resolution is the language pro-drop properties. Another influential factor is the sentence structure, especially the distribution of finite vs. non-finite verb forms. The question of contrastive analysis of anaphoric chains is very interesting. Having been touched upon here they deserve more detailed qualitative and quantitative analysis. Our future work will be to analyze this topic in more detail by addressing separately more extensive parallel and non-parallel texts in the languages.

Acknowledgements

We acknowledge support from the Grant Agency of the Czech Republic (grant P406/12/0658), GAUK 3389/2015, EU (grant FP7-ICT-2013-10-610516—QTLeap) and SVV project number 260 224. On the Russian side, the study was supported by the Russian Foundation for Basic research (grant No. 15-07-09306). This work has been using language resources developed, stored and distributed by the LINDAT/CLARIN project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2010013). We thank the Lomonosov Moscow University students Ru-eval team for Russian data annotation, and Dmitrij Gorshkov for software support (the RuCor creation and Russian data management).

References

1. *Ariel, M.* (2001), Accessibility theory: An overview, in Sanders, T., Schliperoord, J. & W. Spooren (eds.) Text representation: Linguistic and psycholinguistic aspects, Amsterdam, Philadelphia: John Benjamins Publishing, pp. 29–87.
2. *Chomsky N.* (1981), Lectures on government and binding, Dordrecht, 1981.
3. *Finkel J. R., Grenager T., Manning Ch.* (2005), Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling, Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005), pp. 363–370.
4. *Givón T.* (1983), Topic continuity in discourse: introduction, in Topic continuity in discourse: Quantified cross-language studies, Amsterdam.
5. *Hajič J., Hajičová E., Panevová J., Sgall P., Bojar O.j, Cinková S., Fučíková E., Mikulová M., Pajas P., Popelka J., Semecký J., Šindlerová J., Štěpánek J., Toman J., Urešová Z., Žabokrtský Z.* (2012), Announcing Prague Czech-English Dependency Treebank 2.0, Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012), Copyright © European Language Resources Association, İstanbul, Turkey, ISBN 978-2-9517408-7-7, pp. 3153–3160.

6. *Hajičová E., Hladká B., Kučová L.* (2006), An Annotated Corpus as a Test Bed for Discourse Structure Analysis, Proceedings of the Workshop on Constraints in Discourse, Copyright © National University of Ireland, Maynooth, Ireland, pp. 82–89.
7. *Jongejan B., Dalianis H.* (2009), Automatic training of lemmatization rules that handle morphological changes in pre-, in- and suffixes alike. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP.—Singapore: Association for Computational Linguistics, 2009.—P. 145–153.
8. *Kibrik, A. A.* (2011), Reference in discourse, Oxford, Oxford University Press.
9. *Kibrik, A. A.* (2013), Peculiarities and origins of the Russian referential system, in Dik Bakker and Martin Haspelmath (eds.) Languages Across Boundaries: Studies in Memory of Anna Siewierska, Berlin, Mouton de Gruyter.
10. *Kibrik, A. A.* (1997), Modelling of multifactor processes: referential choice in Russian discourse [Modelirovaniye multifaktornogo protsessa: vybor referentsial'nogo sredstva v russkom diskurse], MSU reporter [Vestnik MGU], Vol. 4.
11. *Kunz, K., Degaetano-Ortlieb, S., Lapshinova-Koltunski, E., Menzel, K. and Steiner, E.* (to appear 2015), GECCo—an empirically-based comparison of English-German cohesion, in De Sutter, G. and Delaere, I. and Lefer, M.-A. (eds.). New Ways of Analysing Translational Behaviour in Corpus-Based Translation Studies. TILSM series. Mouton de Gruyter.
12. *Lambrecht, K.* (1994), Information structure and sentence form. Topic, focus and the mental representation of discourse referents, Cambridge, Cambridge University Press.
13. *Linguistic Data Consortium* (1999), Penn Treebank 3. LDC99T42.
14. *Mikulová M., Bémová A., Hajič J., Hajičová E., Havelka J., Kolářová V., Lopatková M., Pajas P., Panevová J., Razímová M., Sgall P., Štěpánek J., Urešová Z., Veselá K., Žabokrtský Z., Kučová L.* (2005), Anotace na tektogramatické rovině Pražského závislostního korpusu. Anotátorská příručka. Technical report no. 2005/TR-2005-28, Copyright © ÚFAL MFF UK, Prague, ISSN 1214-5521, 1185 pp.
15. *Nedoluzhko A., Mírovský J., Fučíková E., Pergler J.* (2014), Annotation of coreference in Prague Czech-English Dependency Treebank. Technical report no. 2014/TR-2014-57, Copyright © ÚFAL MFF UK, ISSN 1214-5521, 41 pp.
16. *Novák M., Nedoluzhko A.* (to appear in 2015), Comparison of coreferential expressions in Czech and English, in Discours, Vol. 16.
17. *Poesio M., Ponzetto, S. Versley, Y.* (2011), Computational models of anaphora resolution: A survey Linguistic Issues in Language Technology, available at http://cswww.essex.ac.uk/poesio/papers.html.
18. *Reuland, Eric J.* (2011), Anaphora and language design, Cambridge, MA: MIT Press.
19. *Salmon-Alt S., Vieira R.* (2002), Nominal Expressions in Multilingual Corpora: Definites and Demonstratives, in Language resources and evaluation conference LREC 2002, Las Palmas, Spain.
20. *Salmon-Alt S., Vieira R., Gasperin C.* (2005), Coreferent and anaphoric demonstrative NPs, in António Branco, Tony McEnery, Ruslan Mitkov (eds.) Anaphora Processing: Linguistic, Cognitive and Computational Modelling, Jonh Benjamons, Lisbon, Portugal.

21. *Schmid H.* (1994), Probabilistic Part-of-Speech Tagging Using Decision Trees // Proceedings of International Conference on New Methods in Language Processing. Manchester. 1994. Vol. 12, Issue 4. P. 44–49.
22. *Sidorova E. V.* (2013), Evolution of subject reference in Russian [Evolutsiya sub'yektnoy referentsii v russkom yazyke], Master thesis.
23. *Straková J., Straka M. and Hajič J.* (2014), Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition, Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pages 13–18, Baltimore, Maryland, June 2014. Association for Computational Linguistics.
24. *Strube, M. and U. Hahn* (1999), Functional Centering: Grounding Referential Coherence in Information Structure. Computational Linguistics 25/3, pp. 309–344.
25. *Testelets, Ya.* (2001), Introduction to general syntax [Vvedenije v obščˇij sintaksis], Moscow: RGGU. 2001.
26. *Toldova, S., Grishina Ju., Ladygina A., Vasilyeva M., Nedoluzhko A., Rojtberg A., Azerkovich I., Kurzukov M., Ivanova A.,* (2014):RU-EVAL-2014: Evaluating Anaphora and Coreference Resolution for Russian. In: Computational Linguistics and Intellectual Technologies,ISSN 2221-7932, 13 (20), pp. 681–694

СТРУКТУРА УСТНОГО ДИСКУРСА: ВЗГЛЯД СО СТОРОНЫ МУЛЬТИМОДАЛЬНОЙ ЛИНГВИСТИКИ[1]

Николаева Ю. В. (julianikk@gmail.com),
Кибрик А. А. (aakibrik@gmail.com),
Федорова О. В. (olga.fedorova@msu.ru)

Институт языкознания РАН
и МГУ имени М. В. Ломоносова, Москва, Россия

Данный доклад представляет собой шаг в направлении мультимодальной лингвистики, рассматривающей вербальную форму устного дискурса наряду с просодическими и жестовыми явлениями, входящими в состав устной коммуникации. Достаточно хорошо известно, что устный дискурс структурируется при помощи просодических характеристик. Базовой единицей речи являются элементарные дискурсивные единицы (ЭДЕ), определяемые при помощи комплекса просодических критериев и коррелирующие с такой семантико-синтаксической единицей как клауза. Единицей более высокого иерархического уровня являются предложения, границы которых также идентифицируются посредством просодических признаков. Иллюстративные жесты также могут сигнализировать о том, какие ЭДЕ говорящий объединяет в предложения. Это осуществляется за счет уподобления жестов по формальным (место производства жеста, конфигурация рук, траектория и направление движения) и содержательным (референты, место и время действия) характеристикам. Один из видов уподобления, удержание, коррелирует с предложением устной речи, а второй, жестовая инерция, с единицей более высокого уровня — эпизодом. Таким образом, мы наблюдаем частичную координацию между различными компонентами мультимодального дискурса.

Ключевые слова: структура дискурса, мультимодальная коммуникация, «Рассказы о грушах», элементарная дискурсивная единица, иллюстративные жесты

[1] Работа выполнена при финансовой поддержке гранта РНФ «Язык как он есть: русский мультимодальный дискурс» (14-18-03819).

DISCOURSE STRUCTURE: A PERSPECTIVE FROM MULTIMODAL LINGUISTICS

Nikolaeva Y. V. (julianikk@gmail.com),
Fedorova O. V. (olga.fedorova@msu.ru),
Kibrik A. A. (aakibrik@gmail.com)

Institute of Linguistics, Russian Academy of Sciences and
Lomonosov Moscow State University, Moscow, Russia

This paper is a step towards multimodal linguistics, considering the verbal form of spoken discourse along with prosodic and gestural phenomena, involved in the process of spoken communication. It is well established that spoken discourse is structured with the help of prosodic features. The basic segment of talk is elementary discourse unit (EDU), defined on the basis of a set of prosodic criteria and correlated with the semantico-syntactic unit known as clause. A hierarchically more complex unit is sentence. Sentence boundaries are also identified by prosodic features. Illustrative gestures can signal EDU combination into sentences, too. This is performed by gesture assimilation in formal (location of gesture, hand configuration, trajectory, and direction of movement) and content-related (referents, place and time of event) characteristics. One kind of gesture assimilation, catchment, correlates with spoken sentence, whereas the other kind, gestural inertia, with a higher level unit, namely episode. We thus observe partial correlation between the components of multimodal discourse.

Keywords: discourse structure, multimodal communication, “Pear stories”, elementary discourse unit, gesticulation

1. Введение

До последнего времени по разным причинам — как объективным, так и субъективным — лингвистика занималась изучением исключительно письменных текстов, оставляя в стороне разнообразные по жанрам образцы устного дискурса. В конце XX в. эта ситуация начала меняться; важной вехой на пути становления анализа устного дискурса стала коллективная монография под ред. У. Чейфа «Рассказы о грушах» (Chafe 1980). В самые последние годы стало очевидно, что реальная коммуникация не исчерпывается передачей информации, содержащейся в вербальной форме — существенная доля приходится на невербальные средства, в том числе просодическую и кинетическую (жестовую) составляющие (Gibbon et al. (eds.) 2000; Kress 2002; Hugot 2007; So et al. 2009;

Loehr 2012; Goldin-Meadow 2014). Говоря другими словами, естественный дискурс по своей природе *мультимодален*[2]. Современная лингвистика все сильнее осознает необходимость в исследовании всех трех основных коммуникативных каналов — вербального, просодического и визуального, подробнее см. (Кибрик 2010; Kress 2010; Knight 2011; Adolphs, Carter 2013; Müller et al. (eds.) 2014; Кибрик, Молчанова 2014); одним из важнейших является вопрос о выделении для каждого канала элементарных единиц, при помощи которых происходит сегментация информации, а также вопрос о координации между каналами в процессе реальной коммуникации. Этому и будет посвящена настоящая статья.

Исследователи письменного дискурса обычно не тратят много времени на установление иерархической структуры текста — каждый пишущий сам делит свой дискурс на клаузы, сложные предложения и абзацы. Совсем иначе устроен устный дискурс, при транскрибировании которого сразу встает вопрос сегментации звукового потока на отдельные составляющие. Базовой единицей локальной структуры устного дискурса является *элементарная дискурсивная единица*, которая выделяется на основании просодических критериев и прототипически соответствует одной клаузе (см. раздел 2). В то же время выделение в устном дискурсе иерархических единиц более высокого уровня — *предложений* и *эпизодов* — сопряжено с бóльшими трудностями. В частности, объем предложения варьирует от 1–2 клауз у одних говорящих до длинных цепочек в 10–15 клауз у других, а само отождествление предложений требует сложной аналитической процедуры; отождествление эпизодов представляется еще более сложной задачей. Таким образом, мы не можем отнести понятия «предложение» и «эпизод» к числу базовых единиц устного дискурса. Тем не менее, на наш взгляд, предложения и эпизоды в устной речи существуют, и подтверждение реальности их существования можно получить при анализе *иллюстративных жестов*, сопровождающих устную речь (Kendon 1994, 2004; McNeill 1995; Николаева 2012, 2013); раздел 3 будет посвящен типологии подобных жестов, как единичных — изобразительных, метадискурсивных, ритмических, указательных и жестовых ударений, так и *жестового уподобления* — жестовых удержаний и жестовой инерции. Наконец, в разделе 4 мы рассмотрим вопрос о *координации* между синтаксическим, просодическим и жестовым уровнями. Мы покажем, что в прототипическом случае единичные жесты оказываются хорошо скоординированы с ЭДЕ, жестовые удержания соответствуют предложениям, а жестовая инерция — эпизодам.

2. Структура устного дискурса: вклад просодии

Устный дискурс порождается не в виде непрерывного потока, а определенными квантами, порциями или толчками; мы называем их *элементарными дискурсивными единицами* (ЭДЕ), подробнее см. (Кибрик 2008; Кибрик, Подлесская (ред.) 2009; Kibrik 2011).

---

[2] Понятие модальности, используемое в психологии, нейрофизиологии и информатике, обозначает канал, посредством которого человек воспринимает внешнюю информацию.

Сегментация устного дискурса на ЭДЕ производится на просодических основаниях. ЭДЕ имеет единый контур частоты основного тона; основной акцентный центр, обычно расположенный на реме; характерный громкостный паттерн: затихание к концу; характерный темповый паттерн: ускорение в начале, замедление к концу; характерный паттерн паузации: пауза сопровождает планирование ЭДЕ и отсутствует внутри ЭДЕ; подробнее см. (Кибрик, Подлесская (ред.) 2009: 57 и сл.)

Прототипические ЭДЕ демонстрируют примечательную координацию физиологических, когнитивных и семантических характеристик: произносятся на одном выдохе, отражают один «фокус сознания» (Chafe 1994), описывают одно событие/состояние. С грамматической точки зрения ЭДЕ представляет собой одну клаузу и чаще всего состоят из 2–4 слов.

Таким образом, мы рассматриваем ЭДЕ как базовую единицу устного дискурса. Обратимся теперь к понятию «*предложение*». Устная речь не содержит пунктуационных знаков, по которым можно было бы определять границы предложений. Какие другие критерии мы можем использовать? Согласно (Кибрик 2008) первый кандидат на эту роль — различие между восходящим и нисходящим несущим акцентом в ЭДЕ. Можно ли утверждать, что каждая ЭДЕ с нисходящим несущим акцентом (=интонация точки) завершает предложение, а ЭДЕ с восходящим несущим акцентом (интонация запятой) является нефинальной? Оказывается, что нет — существуют многочисленные случаи нисходящего акцента, который нельзя охарактеризовать как финальный (подробнее см. (Кибрик 2008: 108 и сл.)). Таким образом, установить границы предложений на основании направления тона оказывается невозможно.

Однако существует два других просодических критерия — критерий целевого уровня движения нисходящего тона и характерное тоновая кривая на постакцентных слогах, помогающих различать финальное и нефинальное падения, подробнее см. (Кибрик 2008: 111 и сл.). Выявление этих критериев позволяет сохранить понятие «предложение» применительно к устному дискурсу. Однако предложения нельзя назвать базовой единицей устного дискурса. Зачем же нужны предложения в устной речи? Возможно, предложения оказываются полезными в качестве промежуточной единицы, которая крупнее ЭДЕ, но мельче эпизода. Это объяснение перекликается с гипотезой Чейфа, согласно которой предложению соответствует «суперфокус сознания» (Chafe 1994: 148).

3. Структура устного дискурса: вклад иллюстративных жестов

Среди мануальных жестов выделяются два больших класса — *жесты*-эмблемы (Efron 1941/1972; Ekman, Friesen 1969) и *иллюстративные жесты* (см. обзор в Kendon 2004). Для первых существует заранее заданная форма и фиксированная связь между означаемым и означающим; вторые, наоборот, носят спонтанный характер, не имеют закрепленной формы и фиксированной связи. Первый класс жестов намного лучше изучен: для них созданы словари (Григорьева и др. 2001; Крейдлин 2002; Акишина, Кано 2010), они изучаются в вузах (Андриенко, Сластенин 2004),

используются в популярной психологии (Кузина 2012). В то же время в реальном общении на два порядка чаще встречаются менее изученные иллюстративные жесты (Николаева 2013). Как показывают исследования (Cassell et al. 1999; Melinger, Levelt 2004; Hostetter 2011; Hall, Knapp (eds.) 2013), иллюстративные жесты не всегда полностью осознаются собеседниками, но при этом существенно влияют на понимание и интерпретацию адресатом обращенной к нему речи.

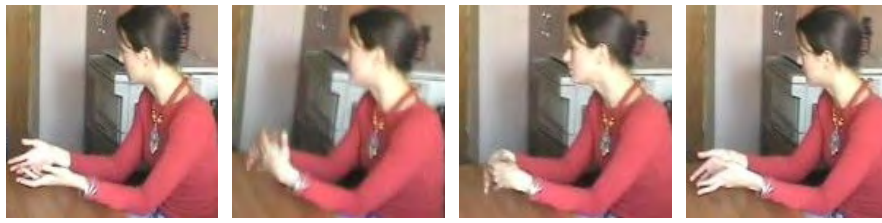
Иллюстративные жесты делятся на следующие типы (упорядочены по частоте встречаемости согласно работе (Николаева 2013)):

- изобразительные жесты (37%), наглядно изображающие черты некоторого референта или действия;
- жестовые ударения (24%), выделяющие некоторый речевой фрагмент аналогично просодии;
- метадискурсивные жесты (16%), отражающие процесс создания дискурса и особенности дискурсивной структуры;
- ритмические жесты (15%), являющиеся средством риторического выделения фрагмента речи;
- указательные жесты (8%), выполняющие непосредственную референцию к присутствующему или воображаемому лицу/объекту.

Жестовыми уподоблениями мы называем серии жестов с повторяющимися характеристиками. Жестовые уподобления выступают в двух формах:

- жестовые удержания (catchment по (McNeill et al. 2001)): одна или более жестовых характеристик (место производства жеста, форма рук и др.) повторяются в серии жестов с сохранением семантики;
- жестовая инерция (inertia по (McNeill et al. 2001)): сохранение внешних характеристик жеста в серии при изменении его семантического наполнения.

На рис. 1 показаны четыре жеста, три из которых сопровождают ЭДЕ 9 и один ЭДЕ 10–11 в примере (1). Жест на рис. 1а изображает обилие груш; жест на рис. 1б — движение к себе, соответствующее глаголу *собирал*, которое показывает действия садовника, кладущего груши в фартук; жест на рис. 1в — движение вниз с грушами, одновременно со словом *спускался*; жест на рис. 1г — движение двумя руками от себя, соответствующее глаголу *выкладывал*. Неизменная конфигурация рук со скругленными пальцами указывает на груши в руках садовника (Николаева 2014). Это случай жестового удержания.

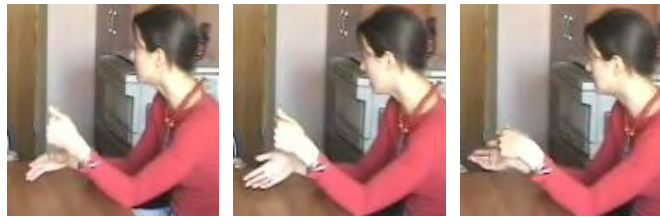


а б в г

Рис. 1. Жестовые удержания

(1)[3]	время, с	№ ЭДЕ	слова и просодия	жесты
	00:16	7	[···(0.5) у него стояло три корзины] с грушами,	изобр.
	00:18	8	и он {[поднимался] на лестницу,	изобр.
	00:20	9	[··(0.3) собирал эти груши] в [ээ(0.3) фартук],	изобр., изобр.
	00:22	10	[··(0.2) спу][скался	изобр.
	00:23	11	и выкладывал]} эти груши в корзину.	изобр.

На рис. 2 показана жестовая инерция. Жест на рис. 2а передает значение внезапной остановки (пример 2: 29)[4]; жест на рис. 2б демонстрирует падение велосипеда (2: 30), на рис. 2в такой же по конфигурации и траектории жест выполнен с бОльшей амплитудой, правая рука в нем показывает упавшую шляпу (2: 31). В этом случае уподобление жестов носит лишь формальный характер, в отличие от случаев жестового удержания, в которых сходные по форме жесты содержат общие семантические элементы.



а б в

Рис. 2. Жестовая инерция

(2)	время, с	№ ЭДЕ	слова и просодия	жесты
	00:59	29	{{[···(0.8) ии эээ(0.8) его велосипед] вре= врезался в камень.	изобр.
	01:02	30	··(0.4) [он] упал,	изобр.
	01:04	31	···(0.7) [с него слетела] шляпа.}}	изобр.

[3] Условные обозначения: многоточие и цифры в скобках обозначают абсолютные паузы, длительность в с; эээ и ммм — заполненные паузы; знак = обозначает обрыв слова, знак == сильный фальстарт; запятая — нефинальность ЭДЕ, точка — финальность ЭДЕ. В квадратных скобках обозначены границы единичных жестов, в фигурных скобках — границы жестовых удержаний, в двойных фигурных скобках — границы жестовой инерции.

[4] Зд. и далее при упоминании примеров первая цифра в скобках обозначает номер примера, цифра после двоеточия — номер ЭДЕ; примеры разделены точкой с запятой.

4. Структура устного дискурса: мультимодальная координация

Ключевым для исследовательской программы мультимодальной лингвистики является вопрос о координации синтаксического, просодического и жестового уровней. В небольшом пока количестве исследований, посвященных этому вопросу (выполненных в парадигме анализа бытового диалога (Ford et al. 2012; Kärkkäinen and Keisanen 2014)), утверждается, что координация достигается далеко не во всех случаях. Мы исследовали этот вопрос на материале 14 пересказов «Фильма о грушах» Чейфа (1980), записанных на видеокамеру и затранскрибированных. Деление на ЭДЕ и предложения было выполнено в программе PRAAT (www.fon.hum.uva.nl/praat) на основании просодических критериев, также была размечена временная динамика и выделены паузы; разметка жестов была произведена в программе ELAN (www.lat-mpi.eu/tools/elan). Необходимым условием работы была независимая разметка вербально-просодических единиц, с одной стороны, и жестов, с другой. В разделе 4.1 мы рассмотрим уровень локальной структуры, в разделе 4.2 — иерархически более высокий уровень предложения, в разделе 4.3 — уровень эпизода.

4.1. Локальная структура дискурса

Прототипическая ЭДЕ содержит один жест, примерно 20% всех ЭДЕ содержат более одного жеста, обычно два; в последнем случае ЭДЕ скоординированы с одиночными жестами одного типа (8: ЭДЕ 71), одиночными разнотипными жестами, а также жестовым уподоблением (1: 9; 3: 18; 5: 68, 69, 72, 73; 7: 71); в редких случаях внутри ЭДЕ может проходить граница между жестами при жестовом удержании (3: 19). В прототипической ЭДЕ жесты по времени «укладываются» в границы ЭДЕ (около 90%), однако для каждого типа жестов имеются свои особенности. Так, изобразительные жесты часто соответствуют целой ЭДЕ (3: 20, 21; 9: 11, 12, 13, 16, 20); жестовые ударения и метадискурсивные жесты чаще других бывают скоординированы с дискурсивными маркерами (4: 114, 118; 5: 70; 9: 17); ритмические жесты часто отмечают целые ЭДЕ; наконец, указательные жесты не обязательно строго совпадают по времени с теми словами, которым они соответствуют семантически (5: 69, 72). Что касается координации жестового уровня и паузации / речевых сбоев, можно отметить, что изобразительные жесты часто начинаются на паузах и речевых сбоях (1: 7, 9, 10; 2: 29, 30, 31; 3: 18; 8: 75), а метадискурсивные жесты и жестовые ударения чаще других отмечают незавершенные или оборванные ЭДЕ (4: 114, 116; 6: 205).

(2)

время, *с*	№ ЭДЕ	вербально-просодический транскрипт	жесты
00:29	17	там {[дерево,]	изобр.
00:30	18	[····(1.2)] к дереву прижата [лестница,]	изобр., изобр.
00:32	19	[и внизу лестницы стоят] [три корзинки,]	изобр., изобр.
00:34	20	[две из которых полные груш,]	изобр.
00:36	21	[а вторая пустая.]}	изобр.

(4)

время, *с*	№ ЭДЕ	вербально-просодический транскрипт	жесты
02:53	113	и он в таких= к= как-то так странно [смотрит на это,]	мета
02:56	114	[вроде как==]	мета
02:57	115	[ну==]	мета
02:57	116	[где==]	мета
02:57	117	[где мои груши,]	мета
02:58	118	и так [расстро]ен видно,	удар.

(5)

время, *с*	№ ЭДЕ	вербально-просодический транскрипт	жесты
02:11	68	{[и ···(0.5)] [что же видит садовник,]	мета
02:13	69	[когда эээ(0.5) ну] [спускается] с [лестницы,]	мета, изобр., указат.
02:15	70	[значит ··(0.2)]	мета
02:16	71	[одной корзинки с грушами нет,]	мета
02:18	72	[··(0.3) а мимо идут три маль][чика	указат.
02:20	73	[и каждый из них ··(0.2) ест грушу,]	мета
02:21	74	ну мм(0.4)]}	

(6)

время, *с*	№ ЭДЕ	вербально-просодический транскрипт	жесты
06:40	204	оди{{[н==	
06:41	205	паца]ны.	удар.
06:42	206	[···(0.6) один==]	указ.
06:43	207	[двое постарше,]}}	удар.
06:44	208	один маленький.	

(7)

время, *с*	№ ЭДЕ	вербально-просодический транскрипт	жесты
01:19	45	потом эээ(0.8) {[следующий кадр],	мета
01:21	46	[это то что проходит] [человек с козой],	мета, мета
01:23	47	[прост= просто мимо проходит человек с козой],	мета
01:25	48	[непонятно почему он там проходит]}.	мета

(8)

время, *с*	№ ЭДЕ	вербально-просодический транскрипт	жесты
02:24	71	···(0.6) эээ(0.6) [эээ(0.7) ···(0.6) эээ(0.8) и вдруг перед ним ··(0.2)] оказываются ··(0.1) несколько ··(0.1) парней,	мета
02:28	72	···(0.6) трое,	
02:29	73	···(0.5) ниоткуда,	
02:30	74	непонятно откуда взявшихся,	
02:31	75	и они {[··(0.2) начинают собирать эти груши,]	изобр.
02:33	76	и [помогать ему складывать]} в корзину.	изобр.

(9)

время, *с*	№ ЭДЕ	вербально-просодический транскрипт	жесты
00:17	8	···(0.7) ну он [приставил==]	изобр.
00:18	9	ну [у него значит лестница к дереву приставлена],	изобр.
00:20	10	[он на нее залезает,]	изобр.
00:21	11	{[собирает груши,]	изобр.
00:22	12	{{[в передничек,]}	изобр.
00:23	13	[··(0.2) с карманом.]	изобр.
00:24	14	потом [спускается,]}}	изобр.
00:25	15	{{[выкладывает их в корзину,	изобр.
00:26	16	так] [любовненько,]	удар.
00:27	17	[причем]}} {[··(0.1) снимает эту свою] [косынку,	удар., изобр.
00:28	18	протирает грушу,]	изобр.
00:29	19	[значит]} ее кладет,	изобр.
00:30	20	[··(0.2) залезает] обратно.	изобр.

В 20 % случаев жесты выходят за границы ЭДЕ, опережая или запаздывая. Можно предположить следующие причины наблюдаемой дискоординации: (i) жест опережает речь в тех случаях, когда моторная программа говорящего, реализующая визуальную иллюстрацию, опережает формулирование высказывания (3: 18; 5: 68); (ii) жест запаздывает, когда говорящий иллюстрирует финальную часть высказывания и не успевает закончить жестикуляцию (9: 16), или же сохранение жеста показывает адресату, что сказанное далее непосредственно связано с той же темой; (iii) рассогласование по времени между словами и жестами может происходить на границе ЭДЕ после сильного фальстарта (6: 205). Важно отметить, что опережающая дискоординация часто сочетается с речевым сбоем или паузой, а запаздывающая может быть связана с особенностями изобразительных жестов, наиболее частотных в данных случаях: они маркируют новую информацию, часто тяготеющую к концу высказывания (3: 17, 18, 19; 9: 10, 14).

4.2. Предложения

Жестовые уподобления — серии жестов, при которых жесты внутри серии связаны между собой более сильными связями, чем обычные единичные жесты. Рассмотрим сначала жестовые удержания, которые могут состоять как из одинаковых по типу единичных жестов (примеры 1, 3, 7, 8), так и из разных (5). Согласно гипотезе МакНила, ЭДЕ, соответствующие жестовым удержаниям, тоже должны быть связаны между собой более сильными связями (McNeill 2000). Предположим, что каждый единичный жест в серии жестовых удержаний соответствует клаузе и просодически заканчивается интонацией запятой, а серия целиком соответствует предложению и маркируется интонацией точки. Проверим это на наших данных, включающих около 150 жестовых удержаний. Оказывается, что все жестовые удержания можно разделить пополам на прототипические и непрототипические случаи. В первом случае каждый жест в серии не выходит за границы ЭДЕ, а границы жестовой серии совпадают с границами предложения (3, 5, 7). Во втором случае жестовая серия оказывается скоординирована с некоторой частью предложения (1, 8). Важно отметить, однако, что границы жестовых удержаний не пересекают границы предложения. Рассмотрев подробнее непрототипические случаи, можно заметить, что они маркируют наиболее значимую (8: 75, 76) часть предложения, а часть других клауз этого предложения сопровождается другими единичными жестами (8: 71, 72).

4.3. Эпизоды

В разделе 2 были описаны просодические критерии выделения ЭДЕ и предложений. Еще более высокий уровень в иерархической структуре дискурса занимают *эпизоды* (van Dijk 1981). Надежные эмпирические методы выделения эпизодов отсутствуют, и по этой причине эпизоды в данной работе были выделены интуитивно на основании семантики. Посмотрим, насколько можно подтвердить предположение о координации эпизодов и жестовой инерции.

В примере (2) можно видеть серию жестов, которые пересекают границу предложения и объединяются в небольшой эпизод. В примере (9) представлен более сложный случай сочетания разных типов жестового уподобления: ЭДЕ (8–10) сопровождаются единичными жестами, затем мы видим жестовое удержание (11–12), на которое частично накладывается жестовая инерция (12–14), а в ЭДЕ (17) также после паузы происходит смена жестовой инерции на жестовое удержание (17–19); наконец, опять же после паузы эпизод заканчивается единичным жестом (20). Таким образом, основываясь на имеющихся данных, можно предположить, что жестовая инерция объединяет еще более крупные иерархические единицы — эпизоды устного дискурса.

5. Заключение

Ключевые единицы устного дискурса определяются на основании просодии: базовым понятием локальной структуры дискурса является ЭДЕ, а предложения

следует рассматривать как второстепенную единицу более высокого уровня. Иллюстративные жесты делятся на более многочисленные единичные жесты и менее многочисленные случаи жестового уподобления. При помощи первых говорящий дополняет и модифицирует свою речь на уровне локальной структуры, при помощи вторых объединяет ЭДЕ в иерархические структуры более высокого уровня — предложения и эпизоды. Прототипическая ЭДЕ скоординирована по времени с одним жестом, прототипические жестовые удержания подчеркивают более тесные связи внутри предложения, а жестовая инерция — внутри целого эпизода.

Данная работа представляет собой шаг на пути создания нового лингвистического направления — *мультимодальной лингвистики*. Общей задачей мультимодальной лингвистики является построение когнитивной модели дискурса, способной непротиворечивым образом объяснять реальное многообразие естественной коммуникации.

Литература

1. *Акишина А. А., Кано Х.* (2010), Словарь русских жестов и мимики, Русский язык, Москва.
2. *Андриенко Е. В., Сластенин В. А.* (2004), Социальная психология: учебник для вузов, «Питер», Санкт-Петербург.
3. *Григорьева С. А., Григорьев Н. В., Крейдлин Г. Е.* (2001), Словарь языка русских жестов, Языки русской культуры, Москва.
4. *Кибрик А. А.* (2008), Есть ли предложение в устной речи, А. В. Архипов, Л. В. Захаров, А. А. Кибрик и др. (ред.), Фонетика и нефонетика. К 70-летию С. В. Кодзасова, ЯСК, Москва, с. 104–115.
5. *Кибрик А. А.* (2010), Мультимодальная лингвистика, Ю. И. Александров, В. Д. Соловьев (ред.), Когнитивные исследования — IV, ИП РАН, Москва, с. 134–152.
6. *Кибрик А. А., Подлесская В. И.* (ред.) (2009), Рассказы о сновидениях: корпусное исследование устного русского дискурса, ЯСК, Москва.
7. *Кибрик А. А., Молчанова Н. Б.* (2014), Каналы мультимодальной коммуникации: относительный вклад в понимание дискурса, О. В. Федорова, А. А. Кибрик (ред.), Мультимодальная коммуникация: теоретические и эмпирические исследования. Сборник статей, Москва, с. 99–114.
8. *Крейдлин Г. Е.* (2002), Невербальная семиотика, Новое литературное обозрение, Москва.
9. *Кузина С.* (2012), Курс начинающего лжеца от А до Я, Астрель, Москва.
10. *Николаева Ю. В.* (2012), Жестикуляция рассказчика и структура нарратива, А. А. Кибрик, Т. В. Черниговская, А. В. Дубасова (ред.), Когнитивные исследования, вып. 5, Институт психологии РАН, Москва.
11. *Николаева Ю. В.* (2013), Иллюстративные жесты в русском дискурсе. Диссертация … кандидата филологических наук, МГУ, Москва.

References

1. *Adolphs S., Carter R.* (2013), Spoken corpus linguistics: From monomodal to multimodal. N.-Y.: Routledge.
2. *Akishina A. A., Kano Kh.* (2010), Dictionary of Russian gestures and mimics [Slovar' russkikh zhestov i mimiki], Russkiy yazyk, Moscow.
3. *Andrienko Ye. V., Slastenin V. A.* (2004), Social psychology: college textbook [Sotsial'naya psikhologiya: uchebnik dlya vuzov], Piter, Saint-Petersbourg.
4. *Cassell J., McNeill D., McCullough K. E.* (1999), Speech-Gesture Mismatches: Evidence for One Underlying Representation of Linguistic and Non-Linguistic Information, Pragmatics and Cognition, vol. 7(1), pp. 1–33.
5. *Chafe W.* (1994), Discourse, consciousness, and time. The flow and displacement of conscious experience in speaking and writing, Chicago.
6. *Chafe W.* (ed.) (1980), The pear stories: Cognitive, cultural, and linguistic aspects of narrative production, Norwood.
7. *van Dijk T.* (1981), Episodes as units of discourse analysis, Deborah Tannen (ed.), Analyzing discourse: Text and talk, Georgetown University Press, Georgetown, pp. 177–195.
8. *Efron, D.* (1941/1972), Gestures, race and culture. The Hague: Mouton.
9. *Ekman P., Friesen W. V.* (1969), The repertoire of nonverbal behavior: Categories, origins, usage, and coding. Semiotica, vol. 1, pp. 49–98.
10. *Ford C. E., Thompson S. A., Drake V.* (2012), Bodily-visual practices and turn continuation, Discourse Processes, Vol. 49(3–4), pp. 192–212.
11. *Gibbon D., Mertins I., Moore R. K.* (eds.) (2000), Handbook of multimodal and spoken dialogue systems: Resources, terminology and product evaluation. Berlin: Springer.
12. *Goldin-Meadow S.* (2014), Widening the lens: What the manual modality reveals about language, learning, and cognition. Philosophical Transactions of the Royal society, vol. 369.
13. *Grigir'yeva S.A., Grigor'yev N.V., Kreydlin G. Ye.* (2001), Dictionary of language of Russian gestures [Slovar' yazyka russkikh zhestov], Yazyki russkoy kul'tury, Moscow.
14. *Hall J. A., Knapp M. L.* (eds.) (2013). Handbooks of communication science: Nonverbal communication, Berlin: De Gruyter Mouton.
15. *Hostetter A. B.* (2011),When do gestures communicate? A meta-analysis. Psychological Bulletin, vol. 137(2), pp. 297–315.
16. *Hugot V.* (2007), Eye gaze analysis in human-human interactions. Master of science thesis. Stockholm, Sweden.
17. *Keisanen T., Kärkkäinen E.* (2014), A multimodal analysis of compliment sequences activity in everyday English interactions, Pragmatics, vol. 24(3), pp. 649–672.
18. *Kendon A.* (1994), Do gestures communicate? A review, Research on Language and Social Interaction, vol. 27, pp. 175–200.
19. *Kendon A.* (2004), Gesture. Visible action as utterance. Cambridge: Cambridge University Press.
20. *Kibrik A. A.* (2011), Cognitive discourse analysis: Local discourse structure, Grygiel M., Janda L. A. (eds.), Slavic linguistics in a cognitive framework, N. Y., pp. 273–304.
21. *Kibrik A. A.* (2008), Is there a sentence in spoken speech [Yest' li predlozhenie v ustnoy rechi], A. V. Archipov, L. V. Zakharov, A. A. Kibrik et al. (eds.), Phonetics

and non-phonetics. Festschrift for 70 of S. V. Kodzasov. [Fonetika i nefonetika. K 70-letiyu S. V. Kodzasova], Languages of Slavic culture, Moscow.
22. *Kibrik A. A.* (2010), Multimodal linguistics [Mul'timodal'naya lingvistika], Yu. I. Aleksandrov, V. D. Solov'yev (eds.), Cognitive studies [Kognitivnyye issledovaniya], Vol. IV, Institute of psychology, Moscow, pp. 134–152.
23. *Kibrik A. A., Podlesskaya V. P.* (eds.) (2009), Nightdream stories: case study of Russian spoken discourse [Rasskazy o snovideniyakh: korpusnoye issledovaniye ustnogo russkogo diskurda], Languages of Slavic culture, Moscow.
24. *Kibrik A. A., Molchanova N. B.* (2014), Channels of multimodal communication: relative input in discourse comprehension [Kanaly mul'timodal'noy kommunikatsii: otnositel'nyy vklad v ponimanie diskursa], O. V. Fedorova, A. A. Kibrik (eds.), Multimodal communication: theoretical and empirical research. Collection of articles [Mul'timodal'naya kommunikatsiya: teoreticheskie i empiricheskie issledovaniya. Sbornik statey], Moscow, pp. 99–114.
25. *Knight D.* (2011), Multimodality and active listenership: A corpus approach. London: Bloomsbury.
26. *Kress G.* (2002), The multimodal landscape of communication, Medien Journal, Vol. 4, pp. 4–19.
27. *Kress G.* (2010), Multimodality: A social semiotic approach to communication, Routledge Falmer, London.
28. *Kreydlin G. E.* (2002), Nonverbal semiotics [Neverbal'naya semiotika], New literary review, Moscow.
29. *Kuzina S.* (2012), Course of beginning liar from A to Z [Kurs nachinayushchego lzhetsa ot A do Ya], Astrel', Moscow.
30. *Loehr D.* (2012). Temporal, structural, and pragmatic synchrony between intonation and gesture, Laboratory Phonology, vol. 3(1), pp. 71–89.
31. *McNeill D.* (1995), Hand and Mind: What Gestures Reveal About Thought, University of Chicago Press, Chicago.
32. *McNeill D., Quek F., McCullough K.-E., Duncan S., Furuyama N., Bryll R., Ma X.-F., Ansari R.* (2001), Catchments, prosody, and discourse, gesture, Vol. 1, pp. 9–33.
33. *McNeill D.* (2000), Growth Points, Catchments, and Contexts. Japanese Journal of Cognitive Science, special issue on gesture, S. Kita (ed.), vol. 7, pp. 22–36.
34. *Melinger A., Levelt W. J. M.* (2004), Gesture and the communicative intention of the speaker, Gesture, vol. 4, pp. 119–141.
35. *Müller C., Fricke E., Cienki A., McNeill D.* (eds.) (2014), Body — Language — Communication. 2 vols., Mouton de Gruyter, Berlin.
36. *Nikolaeva Yu. V.* (2012), Gesticulation of the speaker and discourse structure [Zhestiluyatsiya rasskazchika i sruktura narrativa], A. A. Kibrik, T. V. Chernigovskaya, A. V. Dubasova (eds.) Cognitive studies [Kognitivnyye issledovaniya], vol. 5, Institute of psychology, Moscow.
37. *Nikolaeva Yu. V.* (2013), Gesticulation in Russian discourse [Illustrativnyye zhesty v russkom diskurse]. Diss. cand. philol. science, MSU, Moscow.
38. *So W. C., Kita S., Goldin-Meadow S.* (2009), Using the hands to identify who does what to whom: Gesture and speech go hand-in-hand, Cognitive Science, vol. 33, pp. 115–125.

ГЛАГОЛЫ *БЫТЬ* И *БЫВАТЬ*: ИСТОРИЯ И СОВРЕМЕННОСТЬ[1]

Падучева Е. В. (elena.paducheva@yandex.ru)

ВИНИТИ РАН; МГГУ им. М. А. Шолохова, Москва, Россия

Глагол *бывать* принадлежит к классу глаголов **многократного способа действия**, к которому относятся, например, многократные глаголы несов. вида *хаживать, слыхивать*, соотнесенные с глаголами несовершенного же вида *ходить, слышать*. Однако *бывать* занимает особое место в этом классе. В частности, он единственный из многократных глаголов имеет свойственную глаголам НСВ аналитическую форму **будущего времени**. В работе показано, что эта форма есть у *бывать* не во всех значениях, а только в **переместительном**, ср. допустимое *Я буду бывать у вас чаще,* но **Со временем не будет бывать таких случаев*. Дается объяснение этому факту: глагол *бывать* в переместительном значении входит в аспектуальную систему русского языка не как **итератив** от глагола НСВ *быть*, а как **имперфектив** от глагола СВ *побывать*. Поэтому переместительное *бывать* обладает всеми свойствами имперфектива от моментального глагола: полным набором временных форм, многократным и общефактическим значением несовершенного вида и проч.

Рассматривается соотношение между глаголом *бывать* и глаголом *быть*, про который ранее было выяснено, что он, будучи в своем основном значении имперфективом, может употребляться также в функции глагола СВ, т. е. является биаспектуальным. Обращено внимание на использование *бывать* в качестве **экспрессивного** варианта *быть*.

Ключевые слова: многократность, имперфективация, экзистенциальное значение несовершенного вида, биаспектуальность, экспрессивность

VERBS *BYT'* AND *BYVAT'*: CONTEMPORARY STATE AND HISTORY

Paducheva E. V. (elena.paducheva@yandex.ru)

Sholokhov Moscow State University for the Humanities, Moscow, Russia

The verb BYVAT' 'to be <iteratively>' (formed with the help of an iterative suffix *-yva-* from *byt'*) belongs to the class of verbs of the **iterative Aktionsart**, which includes such verbs as *xazhivat', slyxivat'* related to the imperfective *xodit', slyshat'*. But BYVAT' occupies a special place in that class.

[1] Статья написана при поддержке гранта РНФ 14-18-03270.

In particular, it is the only one to have an analytical form of the **future tense**. It is claimed that this form exists only in the context of the **motional meaning** of BYVAT' , cf. acceptable *Ja budu byvat' u vas chashche* 'I shall be BYVAT' at your place more often' but **So vremenem ne budet byvat' takix sluchaev* 'over time there won't be BYVAT' in such cases'. The following explanation is given to this fact: motional BYVAT' is included in the aspectual system of Russian not as an **iterative** of the imperfective *byt'* but as an **imperfective** of the momentary perfective *pobyvat'*. No wonder that motional BYVAT' possesses all the properties of an imperfective of a momentary verb: a complete set of tense forms, iterative and general factual meaning of aspect, etc.

The relationship is considered between the verb BYVAT' and the verb *byt'*, which was earlier proved to be perfective in some of its uses and, thus, be-aspectual. The attention is drawn to the fact that BYVAT' is often used as an expressive correlate of *byt'*.

Key words: iterativity, imperfectivization, existential meaning of the imperfective, bi-aspectuality, expressivity

1. Постановка задачи

Глаголы *быть* и *бывать* современные словари русского языка трактуют как разные слова. Между тем на протяжении истории они соотносились как глаголы, связанные по виду. Грамматическая традиция XIX — начала XX века трактует образования на *-ыва/-ива-* от основ несов. вида как особый **многократный вид**. Так, в «Русской грамматике» А. Г. Востокова с глаголом *читать*, несовершенного вида, соотносится глагол *прочитать*, совершенного вида, и *читывать* — многократного, см. обзор на эту тему в Виноградов 1947: 119.

Производные на *-ыва/-ива-* (и на *-а-*, *-ва-*) от глаголов несов. вида были необычайно продуктивны в языке XVI–XVIII века (см. приведенные в Успенский 2004 формы *любливать, хачивать, учивать(ся), брасывать, писывать, игрывать,* и многие-многие другие). Позднее образования этого рода, в основном, выпали из литературного языка, хотя до сих пор широко распространены в севернорусских говорах (см. Пожарицкая 1991).

В современном литературном русском языке остался небольшой класс глаголов **многократного способа действия**. Это, например, глаголы *хаживать, слыхивать*, соотнесенные с глаголами несов. вида *ходить, слышать*. Согласно Грамматика 1980: 600, они имеют особую стилистическую окраску — в частности, служат «средством архаизации стиля».

В семантике глаголов типа *хаживать, слыхивать* сочетаются два компонента:

а) многократность (= итеративность);

б) «плюсквамперфектность», т. е. отнесенность к прошлому: в наст. и буд. времени эти глаголы не употребляются.

В этой связи представляет интерес работа Шевелева 2012, где употребление итеративных глаголов типа *хаживать* в русском языке XVI–XVIII века рассматривается на фоне глагола *бывати* — образованного, с помощью того же суффикса *-ыва-*, от глагола *быти*. Важный факт, отмеченный в этой работе,

состоит в том, что глагол *быти*, который является мотивирующим для *бывати*, еще оставался в это время аспектуально неохарактеризованным: он мог употребляться как в имперфективных, так и в перфективных контекстах, т. е. имел не только имперфективное, но и перфективное значение.

Это существенно меняет дело, поскольку от основы совершенного вида суффикс *-ыва/-ива-* мог давать и нормальные имперфективные дериваты, с полным набором значений типологически предусмотренных для имперфективов: в этом наборе многократные значения составляют лишь часть, см. об имперфективном кластере и линейной vs количественной аспектуальности в Плунгян 2011: 408. В таком случае, мы можем считать, что в прошлом глагол *бывати* соотносился с *быти* не как итератив, а как имперфектив (или не только как итератив, но и как имперфектив), и рассматривать соотношение между *быть* и *бывать* в современном языке с этой точки зрения.

Если многократность — это одно из значений несов. вида, то становится понятна связь между многократностью и «плюсквамперфектностью» в семантике глаголов на *-ыва/-ива-* в языке XVI–XVII, о которой идет речь в Успенский 2004, см. ниже пример (2). Сочетание компонентов «неопределенная кратность», «отнесенность к прошлому» и «локализация в сверхдолгих интервалах» (Падучева 2004: 38) типично для несов. вида в современном русском языке — в его общефактическом значении. (Точнее, речь идет об **экзистенциальном** общефактическом значении, см. Падучева 1996: 43–46.) Например, все три компонента входят в семантику несов. вида в предложении (1).

(1) Я *сталкивался* с этой проблемой =
(а) я столкнулся по крайней мере один раз, возможно и больше;
(б) это было в прошлом;
(в) не в ближайшем, а в отдаленном прошлом.

Исходным пунктом для дальнейшего сопоставления *быть* и *бывать* в современном русском языке будет аспектуальная характеристика глагола *быть* как она представлена в Падучева 2013, где демонстрируется биаспектуальность *быть*: показано, что *быть* активно употребляется в современном языке в значении не только несовершенного, но и совершенного вида. Главный новый момент в предлагаемом подходе к аспектуальной характеристике *быть* состоит в том, что аспектуальные различия связываются с различиями в лексической семантике глагола *быть*. (Здесь я во многом опираюсь на описание лексической многозначности глагола *быть* в Апресян 2009: 443–461.) В этом смысле не совсем правильно называть глагол *быть* биаспектуальным: он имеет разную видовую характеристику в разных лексических значениях. Лексические различия играют роль и в морфологии глагола *бывать*.

Итак, обратимся к глаголу *бывать*. В отличие от *быть*, глагол *бывать* однозначно имперфективный. Однако его аспектуальное поведение таит немало сюрпризов. Помимо вида, он имеет, в некоторых значениях, дефектный набор временных форм. Ниже в разделе 2 речь идет об аспектуально-темпоральной уникальности глагола *бывать*. В разделе 3 показано, как глагол *бывать* позволяет расширить наши знания о глаголе *быть*.

2. Вид и время глагола *бывать*

Я исхожу из состава лексических значений, установленных для глагола *бывать* в Активном словаре русского языка под ред. акад. Ю. Д. Апресяна (Словарь 2014).

бывать 1 ‘быть’, связка при именном сказуемом или предикативе, в т. ч. безличном:
Дожди тут *бывают* затяжные; Днем *бывает* жарко.

бывать 2.1 ‘находиться, присутствовать, участвовать’:
Я и не в таких передрягах *бывал*; Он почти не *бывает* дома.

бывать 2.2 ‘время от времени приходить’:
Теперь я стал *бывать* у них часто.

бывать 3.1 ‘существовать, попадаться; входить в поле зрения’:
Бывают щуки в 70 кг весом;
Плохих народов не *бывает*, *бывают* только плохие люди.

бывать 3.2 ‘случаться, неоднократно иметь место’:
Странные *бывают* приключения;
Бывает, что автору трудно расстаться с любимыми персонажами.

В Словаре 2014 имеется еще *бывать* 4; для наших целей это значение можно не отличать от значения *бывать* 3.2.

Глагол *бывать* во всех своих значениях (см., впрочем, одно уточнение ниже) выражает повторяющееся событие или состояние. В ситуации, описываемой глаголом *бывать*, присутствуют три участника: Субъект, Атрибут и Дистрибутор.

Дистрибутор (= распределяющее множество, Падучева 1985: 229–230) — это обстоятельство, которое задает регулярно воспроизводимые отрезки времени (например, *по пятницам*), к которым относится повторяющееся событие/состояние. Дистрибутор может быть замещен показателем кратности (например, *часто*) или количества (например, *дважды*).

В современном русском языке *бывать* употребляется только в контексте повторяющейся (т. е. воспроизводимой) ситуации. Раньше было не так. В языке XVII в. глагол *бывати* мог обозначать длящееся состояние в прошлом, а не повторяющееся событие или положение вещей; см. пример из Успенский 2004: 47, где явно отсутствует многократность:

(2) В Чюдовѣ монастырѣ было отпѣваніѣ по преставльшемся инокѣ схимникѣ Нектаріи, что *бывалъ* Сибирской архиепископъ, да самъ своею волею оставилъ (1667г)

В современном языке такое употребление для бывать не нормативно. Правда, в отрывке (3) из Пушкина *бывало* выражает непрерывно длящееся (издревле) положение вещей — имеется в виду, что всегда будет так, как *было* прежде, а не так, как *бывало* прежде. Но в нем ощущается приятный оттенок архаичности:

(3) Всегда так будет, как *бывало*. Таков издревле белый свет. Ученых много, умных — мало. Знакомых тьма — а друга нет.

Дистрибутор может не иметь поверхностного выражения — *бывать* и само по себе включает идею ‘иногда’, ‘время от времени’:

(4) Он *бывает* несправедлив <иногда>.

Участник **Субъект** в контексте *бывать* 1 может быть и Вещь (т. е. предметная сущность), и Ситуация. Но *бывать* 3.1 и *бывать* 3.2 различаются ровно тем, что у *бывать* 3.1 Субъект Вещь, а у *бывать* 3.2 — Ситуация. У *бывать* 2.1 и 2.2 Субъект может быть только Вещью (обычно это лицо).

Атрибут у *бывать* 1 выражается именным сказуемым или предикативом; у *бывать* 2.1 и 2.2 атрибут — Место.

Лексическая семантика глагола *бывать* как она описана в Словаре 2014 составляет удобную базу для обсуждения аспектуально-временной парадигмы каждого из значений. Вид у *бывать* всегда несовершенный, так что речь может идти только о частных видовых значениях. А парадигма временных форм может быть дефектной.

***бывать* 1**, связка
прош.: Днем *бывало* душно
наст.: Днем *бывает* душно
*буд.: Днем **будет бывать* душно

Как мы видим, у *бывать* 1 есть формы наст. и прош. времени, но нет формы будущего — у *бывать* как глагола НСВ эта форма должна была бы быть аналитической: *будет бывать*. Можно употребить форму простого будущего, *быть*: *Днем будет душно*, но она выражает непрерывно длящееся, а не повторяющееся состояние, которое требуется от *бывать* 1. В контексте однозначного дистрибутива глагол *быть* справляется с выражением итеративного значения в контексте будущего: *По вечерам будет душно*. Но *быть* — это другой глагол, а у *бывать* 1 формы будущего времени нет.

В прошедшем времени *бывать* может быть взаимозаменимо с *быть* в контексте выраженной многократности. Так, в (5) допустимо не только *бывать,* но и *быть*, а в (6) — не только *быть,* но и *бывать*.

(5) Арнольд давал мне множество советов. Но притом никогда *не бывал многословен*. [И. Э. Кио. Иллюзии без иллюзий (1995–1999)]

(6) Наделашин никогда *не был* с заключёнными груб. [А. Солженицын. В круге первом (1968)]

Глагол *бывать* может быть связкой в конструкции со страдательном причастием (но не вспомогательным глаголом в форме страдательного залога):

(7) Когда ни позвонишь, он всегда *бывает* занят.

В контексте (8) лучше сказать не *бывали*, а *были*, поскольку наречие *неизменно* наводит на мысль о непрерывно длящемся состоянии:

(8) Не думаю, что преувеличиваю, поскольку мысли Эмиля Теодоровича неизменно **бывали заняты* цирком. [И. Э. Кио. Иллюзии без иллюзий (1995–1999)]

Инфинитив затруднен; в (9) семантическую нестыковку сглаживает порядок слов, а предложение *Он стал все чаще бывать не в форме* сомнительно:

(9) я всегда находился наготове — на тот случай, если шеф окажется «не в форме». А не в форме он *стал бывать* все чаще… [Михаил Козаков. Актерская книга (1978–1995)]

***бывать* 2.1** ‘находиться, присутствовать, участвовать’
прош.: Я и не в таких передрягах *бывал*.
наст.: Секретарша директора редко *бывает* на месте.
*буд.: —

У *бывать* 2.1, как и у *бывать* 1, нет аналитической формы буд. времени. В (10) *будешь бывать* скорее понимается в значении *бывать* 2.2, чем в 2.1:

(10) Самый лучший вариант распределения. Я же вижу, что тебе до смерти не хочется покидать Питер. А так между рейсами *будешь бывать* здесь. [В. Аксенов. Коллеги (1962)].

Но и наст. время у *бывать* 2.1 под вопросом. Так, в отрицательном контексте *не бывает* лучше заменить на *нет*, т. е. *бывать* — на *быть* (если многократность обеспечивается контекстом):

(11) Каждый день, за исключением тех дней, когда меня *не бывает* дома (= *нет дома*), я закрываюсь у себя в комнате, закладываю бумагу в свою маленькую прожорливую «Колибри» и пишу. [Фазиль Искандер. Начало (1969)]

(12) Когда мамы *не бывает* дома (= *нет дома*), он даже спать уходит к соседке [Коллекция анекдотов: дети (1970–2000)].

В положительном контексте форма наст. времени *быть* нулевая, и употребление *быть* затруднено; но *бывать* кажется не вполне уместным (пример (13) — из МАС):

(13) [Вера] когда *бывает* дома, всегда сидит под портретом госпожи Ельцовой (Тург.)

***бывать* 2.2** ‘приходить’, переместительное значение

Это единственное значение, при котором глагол *бывать* имеет формы всех трех времен.

прош.: Удивительная это была комната! Сколько раз я потом в ней *бывал*, и всегда испытывал какое-то таинственное волнение… [В. Давыдов. Театр моей мечты (2004)]

наст.: Я редко теперь *бываю* в Комарово [А. Городницкий. «И жить еще надежде» (2001)]

буд.: А если она приехала — пусть… вы познакомите ее со мной, *будете бывать* вместе с нею… [З. Н. Гиппиус. Без талисмана (1896)]
Он сказал, что уже больше *не будет бывать* здесь… [А. П. Чехов. Дядя Ваня (1897)]
— Но надеюсь, однако, что вы *будете бывать* у нас. [А. Ф. Писемский. В водовороте (1871)]

Переместительное значение глагол *бывать* имел с давних пор. Пример употребления *бывать* в переместительном значении в языке XV века из Шевелева 2012:

(14) А полковъ князя великого ни един человѣк *не бывалъ* с нимъ за рѣку (НЛ 1492) (= ‘не переправлялся’)

Наличие аналитической формы буд. времени отличает значение *бывать* 2.2 от всех других значений *бывать*. На это следует обратить внимание — казалось бы, *бывать* как глагол несов. вида аналитическую форму будущего времени должен иметь всегда.

Глагол *бывать* во всех значениях, кроме 2.2, несомненный imperfectivum tantum. Таковым он входит в список непарных глаголов несовершенного вида из Маслов 2004: 78, наряду с *жить* или *существовать*. Однако в значении 2.2 глагол *бывать* выступает как парный: он входит в видовую пару, но не с *быть*, а с *побывать*. Толкование *бывать* 2.2 в Словаре 2014 гласит: ‘В моменты времени А3 человек А1 перемещается в место А2, находится там какое-то время, а затем покидает его’. Но моментов времени А3 может не быть. Так, в примере (15) обстоятельство *пару раз* задает не моменты времени, в которые имеет место ситуация ‘бывать’, а указывает, сколько раз происходит **событие**, обозначаемое глаголом СВ *побывать*. Да и там, где моменты времени есть, в каждый из моментов происходит событие (А1 переместился в место А2, находился там какое-то время, а затем покинул его), обозначаемое глаголом СВ *побывать*, а глаголом НСВ описывается вся серия этих событий вместе взятая.

Глагол *побывать* 2.2 моментальный, так что его имперфектив имеет только **тривиальные** значения (см. о тривиальных значениях НСВ Зализняк, Шмелев 46–50). Поэтому лексическое значение *бывать* 2.2 вообще не требует

толкования отдельного от толкования *побывать* 2.2, который обозначает событие. Имея толкование для А1 *побывал в* А2, мы для (15) с глаголом *бывать* 2.2 получаем:

(15) Она *бывала* в этом доме всего пару раз = 'всего пару раз имело место событие: она *побывала* в этом доме'.

Видовая парность *побывать* — *бывать* требует обоснования. Видовая парность проверяется критериями Маслова (см. Маслов 2004: 76–77) — тестом на настоящее историческое и на многократность. Пара *побывать* — *бывать* проходит тест на многократность: *бывал = много раз побывал*. Теста на настоящее историческое эта пара не проходит — потому что повторяющееся событие, обозначаемое глаголом *бывать*, локализуется на сверхдолгих интервалах (о критериях видовой парности см. подробнее в Зализняк, Шмелев 2000: 48).

Поскольку глагол *побывать* моментальный и видовая пара *побывать* — *бывать* тривиальная, она входит в один ряд с парами типа *приходить, находить, случаться, оказываться* и др., см. Зализняк, Шмелев 2000: 56. В тривиальных парах глагол НСВ может употребляться в общефактическом значении или обозначать многократное событие; он не может употребляться в прогрессиве.

Итак, можно сказать, что *бывать* 2.2 имеет форму (и значение) будущего времени потому, что является не итеративом от стативного *быть*, а имперфективом от моментального *побывать*.

Разграничить *бывать* 2.1 и *бывать* 2.2 не всегда просто. В примере (16) Словарь 2014 усматривает *бывать* 2.1. Между тем, скорее это *бывать* 2.2: если бы у *бывать* в (16) было значение 2.1, в предложении было бы генитивное подлежащее.

(16) Он почти *не бывает* дома. Пропадал — летом на Островах, в гребном клубе (И. Грекова).

В примере (17) (из Словарь 2014) тем более значение 2.1, поскольку тут контекст перемещения:

(17) — Гусь, — пояснил он, — редкая здесь птица, *бывает* только пролетом (В. Астафьев).

***бывать* 3.1** 'существовать'
Возможно наст. и прош. время и невозможно будущее.
прош.: Среди них *бывали* настоящие сподвижники.
наст.: *Бывают* на свете добрые люди.
буд.: —

Форма будущего времени отсутствует, как у всех значений *бывать*, кроме 2.2.

Значение существования требует генитивного субъекта при отрицании; здесь *бывать* ведет себя так же, как *быть*:

(18) Русалок *не бывает*.

В значении 3.1, в отличие от всех остальных, глагол *бывать*, по существу, не выражает многократности: положение вещей, описываемое глаголом *бывать* 3.1, нельзя представить как результат итерации никакого однократного. Множественность субъекта в примере *Бывают щуки в 70 кг весом* (из словаря Ушакова) обусловлена тем, что утверждение существования, как правило, осмысленно только в применении к множеству объектов, обладающих определенным свойством (Paducheva 2008). В Словаре 2014 *бывать* 3.1 признано многократным за счет того, что трактуется как ‘попадать в поле зрения, попадаться’ (см. о регулярной многозначности существование /восприятие в Падучева 2004: 150): *попадаться* — очевидный итератив от *попасться*. Однако *попадаться* — это скорее все-таки аналог, чем синоним *существовать*.

***бывать* 3.2** ‘случаться’; ‘многократно или время от времени иметь место на сверхдолгом временном интервале’; о событии, состоянии, положении вещей; подлежащее — предикатное имя или сентенциальный актант.

прош.: *бывало* так, что они по неделям не видались друг с другом. [П. Н. Краснов (1922)]

наст.: Но даже у великих режиссёров *бывают* провалы;
Знаете, *бывает*, что автору и читателю трудно расстаться с любимыми персонажами. [Г. Садулаев (2008)]

*буд.: —

Формы аналитического будущего у *бывать* 3.2 нет. Так, фраза (19) с точки зрения современных норм аномальна — надо употребить не *бывать*, а *быть*:

(19) А со временем это будет самым обыкновенным случаем, а еще со временем и **не будет бывать* других случаев, потому что все люди будут порядочные люди. [Н. Г. Чернышевский (1863)]

Практически исключен инфинитив в контексте (*не*) *может /могло* (*бывать*) — в примере (20) глагол *бывать*, с его подчеркнутой многократностью, кажется избыточным:

(20) В тот момент выяснилось, что величавое одутловатое лицо Ольги Денисовны с застывшей в глазах укоризной *может бывать* добрым и домашним — оно просто нечасто позволяет себе ямочки и улыбки. [Г. Полонский (1975)]

В значении 3.2 глагол *бывать* допускает вводное употребление, в примерах из (21) — в наст. времени, в примерах из (22) — в прошедшем:

(21) Иногда, *бывает*, депрессия накатит. [«Русский репортер», 2009]

В нашей деревне довольно часто отключается электричество и, *бывает*, надолго. [«Наука и жизнь», 2006];

(22) Я их таскал, *бывало*, на плечах, одного на левом, другого на правом. [В. Аксенов (2005)]
Во время войн Минутка переходила из рук в руки, *бывало*, по нескольку раз в день. [«Знамя», 2005]

Заслуживает внимания взаимодействие многократности с отрицанием. Сочетание *никогда не бывало* Р может иметь два значения:
1) событийное — 'ни разу не имело места Р', см. пример (23),
2) стативное — 'ни в какой момент не имело места Р', см. пример (24).

(23) Со мной никогда *не бывало* таких приключений;

(24) Вишь, чего выходит, никогда такого *не бывало*, чтобы лагерь без охраны стоял. [О. Павлов (1993)]

В (24) отрицается существование такого момента времени (в прошлом, сейчас, видимо, не так), на протяжении которого имело бы место состояние 'лагерь стоит без охраны'. Иначе говоря, отрицается, что состояние имело место когда бы то ни было вообще. В (23) отрицается, что событие имело место хотя бы один раз.

Теперь о глаголе *быть*.

3. Глагол *быть* в перфективном употреблении

Глагол *быть* в современном русском языке обычно относят к глаголам imperfectiva tantum, см. Маслов 2004: 78. Между тем, он допускает, наряду с имперфективными употреблениями, также и перфективные, причем в двух разных лексических значениях — в динамическом (переместительном, согласно Апресян 2009), и в стативном, см. примеры (25) и (26) из Падучева 2013:

(25) Врач *был* ровно в семь [= 'прибыл'];

(26) Пока мы приедем, *будет* темно [= 'станет, начнет быть темно'].

Стативное *быть* может быть употреблено в значении СВ только в контексте будущего времени, см. (26'), но переместительное возможно и в буд., и в прош.

(26') Пока мы приехали, **было* (> *стало*) темно.

О наличии у *быть* значения совершенного вида свидетельствует употребление его парного НСВ в «двунаправленном» значении (в данном случае, это

перемещение туда и обратно; в принципе, это значение достигнутого и впоследствии аннулированного результата). В самом деле, двунаправленное значение возможно только у таких глаголов НСВ, которые входят в видовую пару с СВ: глагол СВ задает тот результат, который был достигнут и аннулирован. В данном случае парный СВ — это сам *быть*.

Перфективные значения *быть* описаны в Падучева 2013. В частности, отмечено отсутствие у переместительного *быть* формы наст. времени — в том числе, такой, которая могла бы выступать в узуальном значении НСВ, т. е. в итеративном. Причем не только в утвердительном предложении, где у стативного *быть* эта форма нулевая, но и в отрицательном, где этой формой является *нет*. Этот дефект парадигмы отличает переместительное *быть* от переместительного *бывать*, которое ведет себя как обычный моментальный глагол.

Переместительное *быть* имеет частное видовое значение сов. вида, начинательное, см. (25), и двунаправленное значение несов. вида, неопределенно-кратное, см. (27):

(27) Я *был* в этом доме <возможно, не один раз>.

Переместительное *бывать*, имперфектив, имеет только значение двунаправленного перемещения (многократного). Что понятно: многократность требует возвращения в исходное состояние.

Представляет интерес конкуренция *быть* и *бывать* в контексте отрицания. Так, в (28а) в функции НСВ многократности к переместительному *быть* выступает то же *быть*, и замена *быть* на *бывать* кажется невозможной. Однако в (28б) почти в такой же роли употребляется *бывать*, и этот контекст естественно соотнести с многочисленными другими контекстами, где *бывать* выполняет функцию эмфатического отрицания — усиленного и, возможно, с модальной окраской, см. также (29).

(28) а. Никогда я *не был* на Босфоре. (С. Есенин)
б. На далекой Амазонке *не бывал* я никогда. (Р. Киплинг)

(29) На земле нашей никогда того *не бывало*, чтобы латинская вера была в почете. [Н. И. Костомаров. (1862–1875)].

Почему же итератив *бывать* выступает в роли эмфатического — усиленного — отрицания для *быть*? (Этот вопрос ставится в Шевелева 2012.) Можно думать, дело не в том, что глагол *бывать* выражает множественность. Глагол *быть* тоже выражает множественность и *никогда* в контексте *быть*, как и в контексте *бывать*, имеет **скалярное** значение (Haspelmath 1997: 111–113), т. е. локализует состояние, в том числе, на крайней точке на шкале времени: *никогда не был* = ‘не был, в том числе, в самый отдаленный момент в прошлом’; *никогда не буду* = ‘не буду, в том числе, в самом отделенном будущем’. Дело в том, что *бывать* выражает **избыточную** множественность (ср. Успенский 2004); отсюда экспрессия.

Ту же семантику экспрессивности имеет *бывать* в модальной конструкции с отрицательным инфинитивом: «*не бывать* + Датив» — это экспрессив для ‘не быть’, причем *быть* понимается в совершенном виде буд. времени. Эта конструкция возможна для всех значений *бывать*. Конструкция эгоцентрическая — ее значение предполагает говорящего, что отмечено в Словаре 2014 (не только говорящий уверен в том, что ситуация не наступит; он может сам препятствовать ее наступлению).

- *не бывать* 1 = ‘A1 никогда не станет A3’:

(30) И станцией этой платформе *не бывать* <там точно ничего не реконструируют> (НКРЯ)
Не бывать мне кроткой, послушной женой — была б я сварливая, злая, неугодливая! [П. И. Мельников-Печерский. В лесах. Книга вторая (1871–1874)];

- *не бывать* 3.2 = ‘A1 никогда не произойдет’:

(31) Двум смертям *не бывать*, а одной не миновать;

- *не бывать* 3.1 = ‘A1 никогда не возникнет’:

(32) Москва — третий Рим, а четвертому *не бывать*;

- *не бывать* 2.1/2.2 = ‘A1 никогда не попадет в A3’:

(33) *Не бывать* тебе в Париже.

Экспрессивное отрицание содержит также конструкция *Ничуть не бывало*.

4. Заключение

Итак, обнаружено следующее свойство глагола *бывать*: этот глагол, несомненный имперфектив, при всех значениях, кроме 2.2, не имеет обычной для имперфективных глаголов аналитической формы буд. времени. Предлагаемое объяснение состоит в том, что глагол *бывать*, при всех значениях, кроме 2.2, имеет нехарактерное для русской глагольной системы значение вторичного имперфектива от стативного, т. е. уже имперфективного глагола — каковым является *быть* в его основном значении. Это нетипичный имперфектив, и у него нетипичный набор форм. Между тем в значении 2.2 глагол *бывать* осмысляется как результат имперфективации глагола совершенного вида *побывать*, и поэтому ведет себя и морфологически как обычный имперфектив. Обретя видового партнера СВ, глагол *бывать* получает **событийное** значение. Тем самым он становится обычным **моментальным** глаголом

с полной видо-временной парадигмой, характерной для моментального глагола. А именно, у *бывать* есть:

- будущее время со значением многократно наступающего события в будущем;
- экзистенциальное (так наз. общефактическое) значение события в прошлом с неопределенной кратностью (*бывал в Париже*);
- возможность без ограничений употребляться в инфинитиве (*не мог бывать*).

Отсутствие у *бывать* аналитической формы буд. времени — это следствие его принадлежности к классу глаголов многократного способа действия. В Грамматике 1980: 600 отмечено отсутствие формы буд. времени у глаголов многократного способа действия, но про глагол *бывать* сказано, что он имеет все формы, в том числе, форму будущего — что неточно: *бывать* имеет форму будущего времени только в значении 2.2, где он входит в аспектуальную систему русского языка как парный глагол несов. вида, т. е. как имперфектив от глагола СВ, а не как итератив от НСВ.

Второй факт — ощущение неуместности *бывать* 1 и 2.1 в наст. времени. Часто хочется заменить *бывать* на *быть*: хотя *быть* не выражает итеративности однозначно, но если итеративность выражена контекстом, выражение итеративности в глаголе не необходимо. Так, в примере (34) употребление *бывать* 2.1 по современным нормам избыточно (время прошедшее, но это прошедшее нарративное, которое по смыслу эквивалентно настоящему, поскольку предполагает синхронный ракурс):

(34) Матери Клэр никогда *не бывало* дома [Г. А. Газданов (1930)]

Несов. вид в (34) имеет узуальное значение, которое не нуждается в подчеркивании с помощью вторичного показателя имперфективности. Итеративное *бывать* может быть оправдано разве что в контексте дистрибутора (типа *Когда я приходил*) — которого в контексте нет.

Еще один факт, который может получить объяснение, — это преимущественное употребление *бывать* в прош. времени. О том, что диалектные многократные формы во много раз чаще употребляются в прош. времени, чем в наст. или будущем, см. Пожарицкая 2011; то же отмечается для литературного языка в Грамматика 80: 600. Тут дело в том, что состояния и другие нетерминативные ситуации, обозначаемые имперфективными глаголами (непарными), плохо итерируются. Если речь идет о прошлом, то прошедшее время обеспечивает прекращение состояния — что создает основу для его возобновления. А в настоящем и в будущем этот фактор отсутствует.

Литература

1. *Апресян Ю. Д.* (2009), Исследования по семантике и лексикографии. Т. I. Парадигматика. М., 2009.
2. *Виноградов В. В.* (1947), Русский язык: Грамматическое учение о слове. М.; Л.: Учпедгиз, 1947.
3. *Грамматика* (1980), Русская грамматика. Т. 1–2 / Отв. ред. Н. Ю. Шведова. М.: Наука, 1980
4. *Зализняк Анна А., Шмелев А. Д.* (2000), Лекции по русской аспектологии // М.: Языки рус. культуры, 2000.
5. *Маслов Ю. С.* (2004) Избранные труды. Аспектология. Общее языкознание. М.: ЯСК, 2004.
6. *МАС*, Словарь русского языка. В 4 т. / Ред. А. П. Евгеньева. М.: Рус. яз., 1981.
7. *Падучева Е. В.* (1985) Высказывание и его соотнесенность с действительностью, Издание 6-е, испр. — М.: Изд-во ЛКИ, 2010. http://lexicograph.ruslang.ru/TextPdf1/paducheva1985.pdf
8. *Падучева Е. В.* (1996) Семантические исследования: Семантика времени и вида в русском языке. Семантика нарратива. М.: Языки русской культуры, 1996. http://lexicograph.ruslang.ru/TextPdf1/PaduSemantIssl1996.pdf
9. *Падучева Е. В.* (2004), Динамические модели в семантике лексики. М.: Языки славянской культуры, 2004. http://lexicograph.ruslang.ru/TextPdf1/PaduDinamMod2004.pdf
10. *Падучева Е. В.* (2013), Русский глагол быть: употребления в значении совершенного вида. IV Конференция комиссии по аспектологии Международного комитета славистов. University of Gothenburg, June 10–14, 2013. http://lexicograph.ruslang.ru/TextPdf1/perfective_byt_Gotenborg_fin.pdf
11. *Плунгян В. А.* (2011), Введение в грамматическую семантику: грамматические значения и грамматические системы языков мира. М.: РГГУ, 2011.
12. *Пожарицкая С. К.* (1991), О семантике итеративных глаголов в севернорусских говорах. «Современные русские говоры», М., Наука, 1991, 74–84.
13. *Словарь* (2014), Активный словарь русского языка. Т.1 А-Б. Ответственный редактор — академик Ю. Д. Апресян. М.: Языки славянской культуры, 2014.
14. *Успенский Б. А.* (2004), Часть и целое в русской грамматике. М.: ЯСК, 2004.
15. *Шевелева М. Н.* (2012), Еще раз о бесприставочных итеративах на *-ыва-/-ива-* типа *хаживать* в истории русского языка. РЯНО, 2012, № 1(23), 140–178.
16. *Haspelmath M.* (1997), Indefinite pronouns. Oxford: Clarendon press.
17. *Paducheva E. V.* (2008), Locative and existential meaning of Russian *быть* // Russian linguistics, 2008, Volume 32, Number 3, 147–158.

References

1. *Apresjan Ju. D.* (2009), Issledovanija po semantike i leksikografii. T. I. Paradigmatika. M., 2009.
2. *Grammatika* 1980 — Russkaja grammatika. T. 1–2 / Otv. red. N. Ju. Shvedova. M.: Nauka, 1980
3. *Haspelmath M.* (1997), Indefinite pronouns. Oxford: Clarendon press.
4. *Maslov Ju. S.* (2004) Izbrannye trudy. Aspektologija. Obshhee jazykoznanie. M.: JaSK, 2004.
5. *MAS*, Slovar' russkogo jazyka. V 4 t. / Red. A. P. Evgen'eva. M.: Rus. jaz., 1981.
6. *Paducheva E. V.* (1985) Vyskazyvanie i ego sootnesennost' s dejstvitel'nost'ju, Izdanie 6-e, ispr. — M.: Izd-vo LKI, 2010.
   http://lexicograph.ruslang.ru/TextPdf1/paducheva1985.pdf
7. *Paducheva E. V.* (1996) Semanticheskie issledovanija: Semantika vremeni i vida v russkom jazyke. Semantika narrativa. M.: Jazyki russkoj kul'tury, 1996.
   http://lexicograph.ruslang.ru/TextPdf1/PaduSemantIssl1996.pdf
8. *Paducheva E. V.* (2004), Dinamicheskie modeli v semantike leksiki. M.: Jazyki slavjanskoj kul'tury, 2004.
   http://lexicograph.ruslang.ru/TextPdf1/PaduDinamMod2004.pdf
9. *Paducheva E. V.* (2008), Locative and existential meaning of Russian *быть* // Russian linguistics, 2008, Volume 32, Number 3, 147–158.
10. *Paducheva E. V.* (2013), Russkij glagol byt': upotreblenija v znachenii sovershennogo vida. IV Konferencija komissii po aspektologii Mezhdunarodnogo komiteta slavistov. University of Gothenburg, June 10–14, 2013.
    http://lexicograph.ruslang.ru/TextPdf1/perfective_byt_Gotenborg_fin.pdf
11. *Plungjan V. A.* (2011), Vvedenie v grammaticheskuju semantiku: grammaticheskie znachenija i grammaticheskie sistemy jazykov mira. M.: RGGU, 2011.
12. *Pozharickaja S. K.* (1991), O semantike iterativnyh glagolov v severnorusskih govorah. "Sovremennye russkie govory", M., Nauka, 1991, 74–84.
13. *Sheveleva M. N.* (2012), Eshche raz o bespristavochnyh iterativah na *-yva-/-iva-* tipa hazhivat' v istorii russkogo jazyka. RJaNO, 2012, № 1(23), 140–178.
14. *Slovar'* (2014), Aktivnyj slovar' russkogo jazyka. T. 1 A–B. Otvetstvennyj redaktor — akademik Ju.D.. Apresjan. M.: Jazyki slavjanskoj kul'tury, 2014.
15. *Uspenskij B. A.* (2004), Chast' i celoe v russkoj grammatike. M.: JaSK, 2004.
16. *Vinogradov V. V.* (1947), Russkij jazyk: Grammaticheskoe uchenie o slove. M.; L.: Uchpedgiz, 1947.
17. *Zaliznjak Anna A., Shmelev A. D.* (2000) Lekcii po russkoj aspektologii // M.: Jazyki rus. kul'tury, 2000.

БЫТЬ ИЛИ НЕ БЫТЬ: КОРПУСА КАК ИНДИКАТОРЫ (НЕ)СУЩЕСТВОВАНИЯ

Пиперски А. Ч. (apiperski@gmail.com)

РГГУ / РАНХиГС, Москва, Россия

В статье обсуждаются понятия приемлемости, встречаемости, грамматичности и существования, в первую очередь — связь между корпусной лингвистикой и вопросом о существовании единиц лексикона. Доказывается, что корпуса не могут свидетельствовать о несуществовании слова, поскольку они обычно являются выборками из некоторой генеральной совокупности, а верхняя граница доверительного интервала для частотности на основе выборки всегда больше 0, вне зависимости от частотности, подсчитанной по выборке. Практическое правило таково: если что-то не встретилось в корпусе, оно могло бы встретиться в корпусе того же размера и состава от 0 до 5 раз. Если же единица присутствует в корпусе, это может служить доказательством её существования в языке, но окончательное решение зависит от того, признаем ли мы корпус репрезентирующим ту разновидность языка, которая нас интересует. Таким образом, корпусное исследование не позволяет доказать несуществование, но позволяет доказать существование; однако второй вид доказательства связан с установлением репрезентативности, которое порой влечёт за собой субъективность и оценочность в суждениях.

Ключевые слова: приемлемость, встречаемость, грамматичность, существование, корпусная лингвистика, выборка, генеральная совокупность, доверительный интервал

TO BE OR NOT TO BE: CORPORA AS INDICATORS OF (NON-)EXISTENCE

Piperski A. Ch. (apiperski@gmail.com)

Russian State University for the Humanities / Russian Academy of National Economy and Public Administration, Moscow, Russia

This paper discusses the notions of acceptability, occurrence, grammaticality and existence, and focuses on the relationship between corpus linguistics and the question of the existence of lexical items. Since corpora are almost exclusively samples from larger populations, it is claimed that they cannot provide evidence for non-existence of words, collocations or constructions.

This is because the upper limit of a confidence interval for frequency based on a sample is always greater than zero regardless of the sample frequency. The rule of thumb goes as follows: anything that does not occur in a corpus might have occurred in a similar same-sized corpus zero to five times. If an item occurs in a corpus, this fact can serve as a proof of its existence in the language, but the final decision depends on whether the relevant contexts from the corpus are judged representative of the language variety of interest. In conclusion, I claim that a corpus-based study cannot prove the non-existence of a linguistic item, although it can be used to prove its existence. However, the latter type of proof includes assessing the representativeness of a corpus, which might lead to subjectivity and value judgments.

Keywords: acceptability, occurrence, grammaticality, existence, corpus linguistics, sample, population, confidence interval

1. Introduction: the notions of acceptability, occurrence, grammaticality and existence

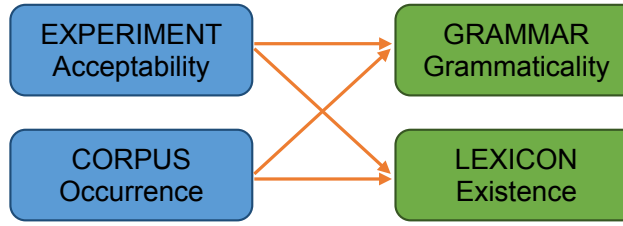
Corpus linguistics has provided linguists with various means of studying frequency-related phenomena. The most radical conceptions of corpus linguistics even state that a corpus is merely a source of information on frequencies (Gries 2009: 11). However, this frequency-based approach is at odds with traditional linguistics which relies heavily on binary distinctions of type: "acceptable vs. unacceptable", "grammatical vs. ungrammatical" and "existent vs. non-existent". Gradient grammaticality has been discussed quite often (cf. Keller 1998, Fanselow et al. (eds.) 2006, Lau et al. 2014; Fedorova 2013 provides a critical survey on this topic and its relation to psycholinguistics), but it is still unwelcome in general linguistics. Scholars are reluctant to accept the idea that grammaticality is gradient rather than categorical, regarding gradience as a matter of performance rather than competence. This raises a question as to whether or not the statistical approach of corpus linguistics generating numerical data and the categorical approach of traditional linguistics can somehow be reconciled. The aim of this paper is to discuss whether statistical data obtained from corpora are transformable into the binary opposition "existence vs. non-existence", which is closely related to grammaticality.

However, in order to do that, we have to make a clear distinction between acceptability, grammaticality and existence, since these terms are sometimes used interchangeably, which might cause confusion. Grammaticality and existence are language-internal, whereas acceptability refers to the speakers' intuitions. As stated by Newmeyer (2007: 398), "no rational linguist would test informants about judgments of grammaticality, since grammaticality is a theoretical construct". In other words, grammaticality is conformity to the rules of the grammar, which cannot be judged without knowledge of these rules. As for existence, it has to do with lexicon rather than grammar. An item is existent if it is listed in the lexicon, and non-existent otherwise. Since lexicon is also a theoretical construct, it is hard to say what kind of items it comprises (cf. Jackendoff 2002 among others), but in general one can say

that grammaticality refers to larger units, e.g., sentences, whereas existence refers to smaller units, e.g., morphemes and words. The position of collocations and constructions on this scale remains debatable.

Corpus frequencies providing a linguist with information about occurrence or non-occurrence are similar to acceptability judgments in that they are both real-life data rather than theoretical constructs. The main source of acceptability data is experiment. I use this term in a broad sense, covering various surveys and tasks as well as introspection, which can be understood as an experiment conducted on a single participant.

Thus, we have two theoretical notions (grammaticality and existence) and two real-life concepts (acceptability and occurrence), and the task of a linguist is to infer information about the former from the latter. This can be summarized in the following scheme:



Scheme 1. The interrelations between acceptability, occurrence, grammaticality and existence

In this paper, I am going to explore only one of the four arrows in Scheme 1, namely the one connecting occurrence and existence. My aim is to answer the following question: Can corpus data tell us whether a word, collocation or construction exists in a given language variety?

2. Absence from corpora as evidence for non-existence?

The absence of an item from a corpus is often taken as evidence of its non-existence. However, this kind of evidence can only be conclusive if a corpus contains the whole population of certain texts rather than a sample. For instance, an absence of a certain word from the Shakespearean canon is sufficient to demonstrate that this word does not occur in the plays of this particular author, but an absence of a word from any corpus of English is not enough to prove that this is not a word of the English language.

Corpus size has a large impact on whether a search returns zero or more results. To illustrate the importance of corpus size, one can compare search results for the same words in two corpora of different sizes. Let us take the main subcorpus of the Russian National Corpus (RNC, www.ruscorpora.ru; 230m words) and the ruTenTen corpus (the.sketchengine.co.uk; 14.5b words), the latter being more than 60 times larger than the former. Table 1 presents a selection of words absent from RNC and their frequencies in ruTenTen:

Table 1: Frequencies of some words absent from RNC in ruTenTen

Word	Absolute frequency	Frequency (ipm)
selfi 'selfie'	4	0.0003
klubneobrazovanie 'tuber formation'	506	0.03
Kuautemok 'Cuauhtémoc (Mexican proper name)'	511	0.03
èkonomist-meždunarodnik 'international economist'	647	0.04
prokrastinacija 'procrastination'	927	0.06
mikruha 'microchip (colloq.)'	2,506	0.17

Clearly, RNC and ruTenTen represent different varieties of Russian, the former including a more standardized language and the language of the 18th, 19th and 20th centuries, but the difference between 0 on the one hand and 506 and 647 on the other hand for such neutral words as *klubneobrazovanie* and *èkonomist-meždunarodnik* is striking. It demonstrates that a word that is absent from a smaller corpus can be quite frequent in a larger corpus. For this reason, the absence of an item from any corpus containing a sample rather than a whole population of some kind cannot be taken as a proof of this item's non-existence in the variety of language represented by this corpus.

3. Confidence intervals instead of binary judgments

Since corpus linguistics is mostly about studying samples, and it is rarely the case that a corpus linguist has to deal with the whole population, standard statistical techniques for estimating population parameters using samples can be applied in this domain. If we are trying to estimate the frequency of a word, we need to construct a confidence interval for the population proportion based on a sample proportion (Baroni & Evert 2009).

Formulae for computing confidence intervals for proportions are given in all basic statistics textbooks (Diez et al. 2012, Field et al. 2012, Rumsey 2010: 77–78, to name just a few recent ones), manuals in statistics for linguists not being an exception (cf. Butler 1985: 62–63, Gries 2013: 129–135). Introductory textbooks usually mention the normal approximation confidence interval:

$$p \pm z\sqrt{\frac{p(1-p)}{n}}$$

where p is the sample proportion, n is the sample size and z is the value of the standard normal distribution corresponding to the desired confidence level (in most cases $z = 1.96$ at the customary confidence level of 95%).

However, this method for computing confidence intervals is inapplicable in the case where p is extremely close or equal to 0 or 1. If $p = 0$, the confidence interval shrinks to [0, 0], which is unsatisfactory: even if we never encounter a phenomenon in our sample, we can never be sure it does not exist at all (cf. Partington 2014). This recalls Laplace's sunrise problem: what is the chance that the sun will rise tomorrow?

Even though the event "the sun does not rise in the morning" has never been observed before, one cannot be sure that its probability is equal to 0. This means that other methods of computing confidence intervals are required.

Fortunately, normal approximation is far from being the only way of estimating confidence intervals for proportions. An extensive list of relevant methods is given in a paper by Newcombe (1998). The most appropriate method for our purposes is Wilson's score method with continuity correction (Newcombe 1998: 859), which is also the default method used by the `prop.test` function in R (R Core Team 2013). The lower limit L and the upper limit U of the confidence interval can be respectively calculated using the following formulae:

$$L = \frac{2np + z^2 - 1 - z\sqrt{z^2 - 2 - \frac{1}{n} + 4p(n(1-p)+1)}}{2(n+z^2)}$$

$$U = \frac{2np + z^2 + 1 + z\sqrt{z^2 + 2 - \frac{1}{n} + 4p(n(1-p)-1)}}{2(n+z^2)}$$

If $p = 0$, the lower limit of the confidence interval L must be taken as 0. If $p = 0$, $z = 1.96$, and we assume that n is much higher than z since sample sizes in corpus linguistics are huge compared to sample sizes in experimental sciences, the expression for U can be simplified:

$$U = \frac{2n \times 0 + z^2 + 1 + z\sqrt{z^2 + 2 - \frac{1}{n} + 4 \times 0 \times (n(1-p)-1)}}{2(n+z^2)} \approx$$

$$\approx \frac{z^2 + 1 + z\sqrt{z^2+2}}{2n} = \frac{4.79}{n}$$

This means that the estimated confidence interval for the proportion given $p = 0$ is $\left[0, \frac{4.79}{n}\right]$. This can be restated as a rule of thumb: if a phenomenon does not occur in a corpus, we can be 95% sure that it will occur 0 to 5 times in a same-sized corpus drawn from the same population.

For this reason, it is not surprising to find a word absent from RNC around 300 times in ruTenTen, which is 60 times larger. This does not even entail the conclusion that RNC and ruTenTen are samples from different populations with respect to the frequency of this particular word. However, the confidence interval approach makes the notion of non-existence virtually non-existent: anything that does not occur in a corpus might have occurred in a similar same-sized corpus 0 to 5 times.

4. Presence in corpora as evidence for existence?

Whether a certain word, collocation or construction is present in a language is often a topic of debate. It is important to answer such questions when writing a text

in the standard variety of a language, composing a dictionary, or creating a grammar. If the judgments of native speakers differ, the presence of an item in a corpus is often quoted as evidence for its existence.

In the Russian-speaking community, it has become increasingly popular to use RNC for investigating the existence of words, collocations or constructions. Since it is the most user-friendly corpus of Russian, accessible even to non-linguists, it is not uncommon to refer to RNC as simply "the corpus", which makes it the sole and ultimate authority for answering questions as to whether something is possible in Russian or not. Discussions of this kind are quite common in social media among educated native speakers, where two positions are clearly identifiable: one group of people might reject some word, collocation or construction because they judge it unacceptable, whereas the other group points to examples drawn from RNC as evidence for the existence of this item.

A typical discussion of this kind took place in 2006 in the blog of the LiveJournal user ormer_fidler(http://ormer-fidler.livejournal.com/22044.html?thread=298268#t298268). A commenter criticized the use of the word *razbudit'sä* instead of *prosnut'sä* 'to wake up' and asked for ormer_fidler's opinion. The latter cited the only example of this word from the RNC as a proof of its existence in Russian, while admitting that this item is very infrequent. Since then, the RNC has experienced a significant increase, and the search now retrieves 5 occurrences of this word. This raises the following question: how many occurrences in a corpus are enough to declare a word existing? Probably the most persuasive answer would be the following: a corpus proves the existence of a word, collocation or construction if it occurs at least once and the retrieved context(s) is (are) judged as relevant to the variety of language in question. The second premise introduces subjectivity into the process of determining what exists in a language and what does not, since contexts from a corpus can be rejected on any grounds, in particular based on value judgments.

A notable feature of RNC is that it incites such value judgments. Since the main subcorpus of RNC contains a high proportion of literary texts (101.8m words / 230m words = 44%), and search results are displayed together with the author's name and the title of the text, users of RNC tend to give more weight to the authors and texts they know and hold in esteem. If a word, collocation or construction was used by a distinguished writer (excluding the writers whose language is unanimously recognized as bizarre, such as Andrei Platonov), users of RNC tend to find it acceptable even if it is very infrequent. However, singular instances can be discarded as "errors" regardless of the status of their author. In other words, if we assume that our corpus is a sample, some parts of it can be claimed to have found their way into the corpus by mistake and not to belong to the population of interest, e.g., "correct standard language". In a recent magazine article, Naberezhnov (2013) provides an instructive example of this kind: when faced with the question of whether *iskrenno sprosit'* 'to ask sincerely' is an acceptable collocation, the organizer of the *Total'nyj diktant* (Total Dictation) project resorts to RNC and finds a single occurrence of this word combination in Boris Pasternak's *Doctor Zhivago*. However, she rejects it as "an unfortunate wording, even though produced by a great writer". Unfortunately, this result is irreproducible. The form *iskrenno* 'sincerely' does not occur all in *Doctor Zhivago*; the

variant form *iskrenne* occurs five times, but it is never used with the word *sprosit'* 'to ask'[1]. Even though unreliable, this example highlights a typical way of using a corpus to prove existence of a collocation.

5. Conclusions

Corpus linguistics is hard to reconcile with the traditional binary distinction "existent vs. non-existent". When doing corpus-based research, linguists need to be more aware of the fact that they are working with samples, which means that they have to apply standard statistical techniques for estimating population parameters from a sample rather than tacitly transfer sample parameters to the whole population. If one bears in mind the nature of a corpus, two conclusions emerge:

a) the absence of a word, collocation or construction from a corpus cannot prove its non-existence, since the upper limit of the confidence interval for its frequency is always above zero.
b) even a single example in a corpus is enough to prove that a word, collocation or construction exists in a language, under the premise that the relevant example(s) can be judged representative of the language variety in question; however, this premise inevitably leads to a certain degree of subjectivity.

References

1. *Baroni, Marco & Stefan Evert.* 2009. Statistical methods for corpus exploitation. In Anke Lüdeling and Merja Kytö (eds.), Corpus linguistics: An international handbook. Vol. 2 (Handbooks of Linguistics and Communication Science 29.2). 777–803.
2. *Butler, Christopher.* 1985. Statistics in linguistics. Oxford: Blackwell.
3. *Diez, David M., Christopher D. Barr, and Mine Çetinkaya-Rundel.* 2012. OpenIntro statistics. [Lexington, KY: CreateSpace].
4. *Fanselow, Gisbert et al.* (eds). 2006. Gradience in grammar: Generative perspectives. Oxford; New York: Oxford University Press.
5. *Fedorova, Olga V.* 2013. Ob èxperimental'nom sintaksise i o sintaksicheskom experimente v jazykoznanii (On experimental syntax and syntactic judgment experiments in linguistics). Voprosy Jazykoznanija 1, 3–21.
6. *Field, Andy P., Jeremy Miles, and Zoë Field.* 2012. Discovering statistics using R. London: Sage.
7. *Gries, Stefan Thomas.* 2009. Quantitative corpus linguistics with R: a practical introduction. New York: Routledge.
8. *Gries, Stefan Thomas.* 2013. Statistics for linguistics with R: A practical introduction: Textbook. Berlin: De Gruyter Mouton.

[1] I am grateful to Vladimir Belikov for pointing this out.

9. *Jackendoff, Ray.* 2002. What's in the lexicon? In S. G. Nooteboom, Fred Weerman and Frank Wijnen. Storage and computation in the language faculty. Dordrecht: Kluwer Academic. 23–58.
10. *Keller, Frank.* 1998. Gradient Grammaticality as an Effect of Selective Constraint Re-ranking In: M. Catherine Gruber, Derrick Higgins, Kenneth S. Olson, and Tamra Wysocki (eds.). Papers from the 34th Meeting of the Chicago Linguistic Society. Vol. 2: The Panels. 95–109.
11. *Lau, Jey Han, Alexander Clark, and Shalom Lappin.* 2014. Measuring gradience in speakers' grammaticality judgements. In: Proceedings of the 36th Annual Meeting of the Cognitive Science Society Québec City, Canada, 23–26 July 2014. 821–826.
12. *Naberezhnov, Grigory.* 2013. Avtorskaja diktatura (Author's dictatorship). Russkij Reporter, 01.04.2013. URL: http://rusrep.ru/article/2013/04/01/totaldiktant
13. *Newcombe, Robert G.* 1998. Two-sided confidence intervals for the single proportion: Comparison of seven methods. Statistics in Medicine 17, 857–872.
14. *Newmeyer, Frederick J.* 2007. Commentary on Sam Featherston, 'Data in generative grammar: The stick and the carrot'. Theoretical Linguistics 33:3, 395–399
15. *Partington, Alan.* 2014. Mind the gaps. International Journal of Corpus Linguistics 19:1, 118–146.
16. *R Core Team.* 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org
17. *Rumsey, Deborah J.* 2010. Statistics essentials for dummies. Indianapolis: Wiley Pub., Inc.

«*И НЕ ДРУГ, И НЕ ВРАГ, А ТАК…*»: ДИСТРИБУЦИЯ И ПРОСОДИЯ МАРКЕРОВ НЕРЕЛЕВАНТНОСТИ ПО ДАННЫМ МУЛЬТИМЕДИЙНОГО КОРПУСА *МУРКО*

Подлесская В. И. (podlesskaya@ocrus.ru)

Российский государственный гуманитарный университет;
Российская академия народного хозяйства
и государственной службы, Москва, Россия

Ключевые слова: нечеткая номинация, русский язык, корпус, устная речь, просодия

"I NE DRUG, I NE VRAG, A TAK…": DISTRIBUTION AND PROSODY OF DISCOURSE MARKERS THAT SIGNAL IRRELEVANCE (EVIDENCE FROM THE MULTIMODAL SUBCORPUS OF THE RUSSIAN NATIONAL CORPUS)

Podlesskaya V. I. (podlesskaya@ocrus.ru)

Russian State University for the Humanities;
Russian Academy of National Economy and Public Administration, Moscow, Russia

The paper focuses on three Russian discourse markers *tak*, *prosto* and *prosto tak*, which fall under a broad category of what is called "loose uses" of language or "vague reference". These are lexical, grammatical and prosodic resources that allow the speaker to refer to objects and events for which the speaker fails to retrieve the exact name, or simply finds the exact name to be unnecessary or inappropriate. The examined discourse markers are employed to signal that the actual state of affairs is less relevant than another (overtly mentioned or implied) one. The three markers are shown to be associated with different information, syntactic and prosodic structures (e. g. pitch movements). The provided qualitative and quantitative analysis is based on data from the multimodal subcorpus of the Russian National corpus.

Key words: vague nomination, Russian, corpus, natural discourse, prosody

1. Постановка вопроса

Данная работа посвящена особому типу нечеткой номинации, при которой говорящий апеллирует к несущественности, незначительности обозначаемого объекта, признака или положения дел[1]. В русском языке нечеткая номинация этого типа представлена, в частности, в конструкциях со словом *так*:

(1) *по делам или так, погулять*…(Б. Окуджава)

(2) *и не друг, и не враг, а так*… (В. Высоцкий)

Для слов типа *так* в указанном значении мы будем использовать ярлык «маркер нерелевантности». В работе на материале мультимедийного корпуса МУРКО будут рассмотрены три маркера такого рода: два однословных — *так* и *просто*, и сочетание *просто так*. Мы постараемся показать, что в живой речи маркеры нерелевантности обладают особыми коммуникативно-просодическими свойствами, а в ряде случаев сопровождаются характерными жестами. В таких работах, как Булыгина, Шмелев (1988, 1997), Падучева (1997), Янко (2008, 2015), и ряде других (см. подробную библиографию в Янко (2015)) было убедительно продемонстрировано, что лексемы могут обладать индивидуальными, словарно закрепленными коммуникативно-просодическими свойствами, например, способностью выступать только в составе рематической или только в составе тематической составляющей, способностью или неспособностью принимать фразовый акцент и рядом других. В редких случаях сведения такого рода фактически отражаются в словарях. Таким счастливым исключением является, например, «Грамматический словарь русского языка», в котором предмет нашего исследования, слово *так* представлено четырьмя вокабулами: «[1]*так* (без удар.) союз; [2]*так* (без удар.) част. (*так что же делать?, так ты не веришь?, вот так штука, так вот, так нет* и т. п.); [1]*так* н.; [2]*так* част. (*что с тобой — так, ничего; он невежда: так, он не знает даже что*…), см. (Зализняк 1980: 440–441). Как видим, интересующая нас функция маркера нерелевантности закреплена в словаре за ударным вариантом частицы (четвертая из перечисленных вокабул). Возникшие в недавнее время корпуса устной речи позволили перейти к изучению просодических приоритетов лексем в естественном дискурсе. Так, например, корпусные данные свидетельствуют, что индивидуальными коммуникативно-просодическими свойствами обладают маркеры, регулирующие развертывание дискурса, типа *ну* или *вот*; например, они обладают разной способностью к автономизации, т. е. разной способностью образовывать отдельную, просодически цельнооформленную реплику (см. подробнее, Подлесская, Кибрик 2009). Развивая эти подходы, в данной работе мы исследуем типы коммуникативно-просодических составляющих, ассоциированных с маркерами нерелевантности, типы реализуемых на них тональных акцентов и ряд других свойств, отличающих их от других лексем, и от других вокабул той же лексемы.

[1] Работа поддержана РФФИ (грант №13-06-00179)

2. Маркер нерелевантности *так*

Для того, чтобы оценить, насколько широко — по данным МУРКО — представлено значение нерелевантности в общей структуре употребления *так* и какие просодические конфигурации ассоциированы с этим значением, было проанализировано сто видеофрагментов в случайной выборке из имеющихся в корпусе 29 279 вхождений по запросу «*так*» после ручного отсева фразеологизированных комплексов *и так далее*, *так тебе и надо*, *так сказать*, *как же так* и им подобных. В транскриптах всех обследованных фрагментов было размечено расположение фразовых акцентов. Как и следовало ожидать, максимальное число вхождений (48 из 100) приходится на наречное употребление *так* в значении способа действия или степени. Это употребление допускает как ударную, ср. (3), так и безударную, ср. (4), реализацию, которые распределены, практически, поровну и систематически следуют общим правилам выбора акцентоносителя, см. Янко 2008. Здесь и далее в работе примеры из МУРКО даются в аутентичном графическом виде, к которому мы добавляем шрифтовое выделение лексем-акцентоносителей с помощью малых прописных букв:

(3) Краса́вица ты́ моя́ / ты́ ж у меня́ загляде́нье! Зна́ешь / е́сли б я́ така́я была́ / да ра́зве я́ ТАК провела́ свою́ жи́знь?

(4) Вы́ счита́ете / что о́н в чём-то винова́т? Да не́т. Э́то ОН та́к счита́ет.

Следующая по численности группа вхождений (23 из 100) объединяет употребления *так* в качестве коннектора. Это могут быть собственно союзные употребления в аподозисе импликативных конструкций, (5), и употребления в качестве дискурсивного коннектора с импликативным значением, отсылающим не к структурно очерченному компоненту полипредикативной конструкции, а к более широкому предтексту, (6). Во всех употреблениях этой группы *так* является безударным:

(5) Зна́чит / я́ то́же неве́жливый? — Бы́л бы ВЕ́ЖЛИВЫМ / та́к давно́ б у меня́ КОРЗИ́НКУ взя́л

(6) Ты́ что́ / в постано́вках игра́ла? В самоде́ятельной худо́жественности? — Угу́. — Ну и что́ / ка́к? Го́лос е́сть / не́т? Что́ говоря́т-то? — Да я́ не зна́ю. Мне́ почётную гра́моту одна́жды присуди́ли. — Так у меня́ ЗНАКО́МЫЙ е́сть по э́той ча́сти / челове́к иску́сства. В Теа́тре рабо́тает / опере́тты.

Напротив, всегда ударными являются *так* в следующей группе (19 из 100) употреблений, которые можно объединить в разряд интерактивных. Сюда входят интродуктивные реплики, реактивные реплики согласия/подтверждения, восклицательные или вопросительные реплики-комментарии. Во всех этих случаях *так* является акцентоносителем, часто реализуется с эмфазой, проявляющейся в удлинении ударной гласной и сложным (восходяще-нисходящим или нисходяще-восходящим движением) тона:

(7) Ки́ра Анато́льна / я́ ва́с о́чень прошу́! Пожа́луйста / у меня́ бо́льше не́т си́л… и вре́мени! Ну́… посмотри́те на часы́! — ТА-АК! Вы́ ИЗДЕВА́ЕТЕСЬ надо мно́й!

Один пример в обследованной выборке демонстрирует употребление *так* в качестве аппроксиматора в составе количественной конструкции (см. подробнее Подлесская 2013), в этой функции *так* всегда безударно:

(8) По́сле у́жина / так ча́сиков в два́дцать но́ль-но́ль. О́чень про́шу!

И, наконец, переходим непосредственно к тем случаям, где *так* выступает в качестве маркера нерелевантности. Их немного, всего 9 из 100, и во всех этих случаях *так* является акцентоносителем ремы, формируя предикацию либо изолированно, (9)–(10), либо в сопровождении минимального набора спецификаторов (11)–(13); крайне редко (единственный пример в нашей выборке) *так* может маркировать нерелевантность в составе контрастной рематической составляющей, имеющей адвербиальный статус, (14):

(9) А о чём молча́ть? — ТА́К / на вся́кий слу́чай. Вдру́г разоблачи́шь. Ма́та Ха́ри.

(10) О́й. А она́ не из Ки́ева? — По-мо́ему / да́. А что́? — Да ТА́К / НИЧЕГО́…

(11) Старики́! Мне́ ди́ск ну́жен. «Пи́нк Фло́йд», «The Dark Side of the Moon». Не помо́жете? — Ты́ чё, па́рень? Мы́ ту́т ТА́К, све́жим ВО́ЗДУХОМ ды́шим. Каки́е ди́ски?

(12) А́х / так вы́ наме́рены но́мер заня́ть? — Да́. Коне́чно. — А я́ уж было поду́мал / что вы́ ТА́К / пря́мо ко МНЕ́.

(13) А́х / го́споди бо́же мо́й! Су́д уже́ бы́л. Э́тим дву́м да́ли по четы́ре го́да. Но э́то ТА́К / МЕ́ЛОЧЬ. А́… э́тому / с кото́рым я́ ва́с спу́тал / во́семь ле́т.

(14) Прика́жете везти́? — Да́ / в Ноттингхи́лл. Де́ньги забы́л. Ох… — Ничего́ / я́ могу́ ТА́К довезти́.

Приведенные примеры показывают, что сам маркер не отсылает к нерелевантному признаку/положению дел, а скорее, сигнализирует о том, что подразумевавшееся релевантное положение дел не имеет места. Это отсутствие часто подчеркивается в составе следующего за маркером парцеллированного уточнения (*ничего*, *мелочь* и т. п.). Иногда релевантное положение дел непосредственно называется в тексте, например, в составе конструкций альтернативного выбора, ср. *по делам или так, погулять* в (1). Однако гораздо чаще релевантное положение дел прямо не эксплицируется, но выводится из более широкого контекста. Так, (11) — это диалог потенциального покупателя с фарцовщиком, который отрицает релевантный для локуторов факт допустимости

купли-продажи («мы не торгуем»). В (12) отрицается ошибочно предполагавшееся намерение князя Мышкина поселиться у Епанчиных. В (13) срок заключения (четыре года) объявляется недостаточно продолжительным, чтобы квалифицироваться как существенный. В (14) релевантным фактором является наличие денег, а *так* означает «без денег». В диалоге *так* часто указывает на отсутствие релевантной причины события или иллокутивной причины, например, отсутствие существенного мотива для вопроса (*А что? — Так, ничего*).

Итак, среди различных значений слова *так*, маркирование нерелевантности встречается достаточно редко, менее чем в 10% случаев в обследованной выборке. Во всех случаях такого рода *так* является акцентоносителем ремы, иногда контрастной. В следующем разделе мы увидим, что маркер нерелевантности *просто*, несмотря на функциональную близость с *так*, имеет другую дистрибуцию и другие просодические свойства.

3. Маркер нерелевантности *просто*

Прежде всего, отметим, что для *просто*, в отличие от *так*, сигнализировать о том, что подразумевавшееся релевантное положение дел не имеет места, это главная функция. Как и в случае с *так*, мы проанализировали сто видеофрагментов в случайной выборке из имеющихся в корпусе 4 768 вхождений по запросу «*просто*»; в транскриптах всех обследованных фрагментов было размечено расположение фразовых акцентов.

Выяснилось, что в выборке всего пять вхождений из ста — это употребления *просто* в качестве наречия/предикатива; во всех этих случаях *просто* является акцентоносителем:

(15) вопро́с о любви́ реша́ется ПРО́СТО.

(16) И Пётр Ильи́ч изба́вится от э́того несча́стья? — Разуме́ется. — Бо́же / ка́к ПРО́СТО.

Во всех остальных случаях *просто* ведет себя как частица, маркирующая нерелевантность. Прототипически эта частица является безударной (72 случая из 100):

(17) Мо́жет быть / вы́ два́дцать ле́т про́сто моро́чили пу́блике ГО́ЛОВУ / а́?

(18) Не на́до ничё обеща́ть. — Я́ не ОБЕЩА́Ю. Про́сто говорю́ / что́ БУ́ДЕТ

Для того чтобы акцент в просодической составляющей сместился на частицу, должны возникнуть специальные условия. К ним относится, прежде всего, контраст — при экспилицитном или имплицитном противопоставлении. В имеющихся восьми случаях такого рода в выборке в контексте контраста частица становится единственным акцентоносителем в составляющей:

(19) А я́ не ПРО́СТО же́нщина. Я́ певи́ца.

(20) Ла́дно / вре́мя пока́ ТЕ́РПИТ. — Почему́ «ПОКА́»? Вре́мя ПРО́СТО те́рпит. — Вре́мя ПОКА́ те́рпит.

Другим фактором, приводящим к ударности частицы, является эмфаза. В семи случаях такого рода в нашей выборке эмфаза приводит к двухакцентной конфигурации типа «шляпа» — на акцентированной частице реализуется подъем тона, а на слове, которое в отсутствии эмфазы являлось бы единственным акцентоносителем ремы, реализуется падение тона:

(21) Мы́ ПРО́СТО вы́пьем ПИ́ВА!

(22) Е́сть у кого́ попроси́ть ПРО́СТО стака́н ВОДЫ́.

Наконец, акцент может возникнуть на частице в составе конструкций с парцелляцией. В таких конструкциях (в нашей выборке их обнаружено восемь) дискурс разворачивается эшелонировано, дробными фрагментами, каждый из которых оформляется как самостоятельная рема. В приведенных ниже примерах и на самой частице, и на следующем за ней фрагменте реализуется типичное для ремы падение тона:

(23) ПРО́СТО/ ради ИНТЕРЕ́СА/ ско́лько ва́м предложи́л аге́нт изра́ильской развёдки Ме́льдман за изме́ну?

(24) А друго́й мо́г бы ПРО́СТО /ОБЧИ́СТИТЬ!

Конструкции с парцелляцией практически смыкаются с конструкциями, в которых *просто* выполняет функцию предикатива. И в случае с парцелляцией, и, как было отмечено выше, в случае предикативного употребления на *просто* размещается фразовый акцент. Близость этих функций хорошо заметна в примерах с парцелляцией, где в качестве параметра (не)релевантности выступает степень вежливости при обращении («с отчеством / титулом vs. без отчества / титула»):

(25) Ка́тя. — Анто́н. — А по о́тчеству? — Для ва́с / ПРО́СТО / АНТО́Н. О́тчество я́ приберега́ю для подчинённых.

(26) Зови́те меня́ ПРО́СТО / СНО́РК.

Фактически, в (25) и, особенно, в (26) *просто* можно считать наречием / предикативом, формирующим рему, за которой следует парцеллированное уточнение. Любопытно, что в нашей выборке имеются и примеры с тем же параметром (не)релевантности (вежливость при обращении), но без парцелляции — с безударной частицей, (27), (28), или с контрастным акцентом на частице, (29):

(27) Серге́й НИКОЛА́ИЧ. Для ва́С / про́сто СЕРЁЖА.

(28) РЕБЯ́ТА / называ́йте меня́ про́сто КО́ЛЯ

(29) Скажи́те / Зо́я-я…[ожидает подсказки отчества, *В.И.П.*] — ПРО́СТО Зо́я.

Наиболее интересной особенностью частицы *просто* в функции маркера нерелевантности является следующая: независимо от того, реализуется ли частица в безударном или в ударном варианте, она всегда водит в состав рематической составляющей. Рематическая составляющая при этом может иметь разный синтаксический статус. Чаще всего — это глагольная группа, просто потому, что именно глагольные группы чаще формируют рему:

(30) Ку́зя / переста́нь му́чить соба́ку! Слы́шишь? — А я́ никого́ не му́чаю! Я́ про́сто де́лаю СЛОНА́!

Но возможно употребление *просто* и в составе именной группы, формирующей рему — в предикативной позиции, (31) или даже в позиции актанта (32).

(31) Вы́ про́сто МОЛОДЦЫ́.

(32) е́сли вещество́ остана́вливает програ́мму старе́ния/ то оно́ должно́ де́йствовать не на како́й-то оди́н ви́д…но та́кже и на други́е ви́ды/ кото́рые старе́ют… Оно́ должно́ лечи́ть не про́сто каку́ю-то одну́ ста́рческую БОЛЕ́ЗНЬ/ допу́стим/ та́м/ КАТАРА́КТА.

Правило о том, что *просто* входит в состав рематической составляющей, не выполняется только в тех случаях, когда *просто* выступает как дискурсивный коннектор (в частном случае, союз) — сохраняя при этом компонент значения, связанный с нерелевантностью. Таких примеров в нашей выборке девять, и во всех *просто* не несет на себе акцента. При этом *просто* располагается преимущественно в крайней левой позиции, см. (33), (34), однако в редких случаях может сдвигаться и внутрь своей сферы действия, см. (35):

(33) Извини́те / что я́ ва́с та́к огорчи́л. — О́ / ничего́. Про́сто у на́с ДО́МА / Но́белевская ПРЕ́МИЯ / запрещённая ТЕ́МА. — Про́сто у на́с ДО́МА / мсьё Гу́ров / о́чень мно́го СТРА́ННОСТЕЙ.

(34) Профе́ссор про́сит ва́с никуда́ не уходи́ть из кварти́ры. Э́то не от недове́рия к ва́м / про́сто кто́-нибудь ПРИДЁТ / а вы́ не ВЫ́ДЕРЖИТЕ и ОТКРО́ЕТЕ.

(35) Не́т / вы́ молоде́ц! Молоде́ц! Вы́ зна́ете / у ва́с таки́е ра́зные весовы́е катего́рии / ма́сса / а вы́ не побоя́лись. . — Не́т / ну что́ вы́ / у меня́ про́сто в ЧЕМОДА́НЕ / случа́йно оказа́лись ГАНТЕ́ЛИ.

Как видно из примеров (33)–(35), несмотря на то, что семантической сферой действия дискурсивного коннектора *просто* является всё следующее за ним предложение, просодически оно примыкает к (или встраивается внутрь) первой коммуникативно-просодической составляющей, которая часто оказывается тематической, ср. *просто у нас дома* в (33), *просто кто-нибудь придет* в (34), *у меня просто в чемодане* в (35).

Как и в случае с *так*, само дискурсивное слово *просто* не отсылает к нерелевантному признаку/положению дел, а сигнализирует о том, что подразумевавшееся релевантное положение дел не имеет места. При этом, по сравнению с *так*, оно дополнительно обладает способностью отсылать к шкале некоторого признака/параметра, указывая на то, что актуально реализованная степень его проявления отличается от предполагавшейся. При этом предполагавшееся значение признака/параметра может выводиться из широкого контекста, а может эксплицироваться в тексте, например, через различные формулы отрицания — «*не X, (а) просто У*», (36), конструкции с предлогом *без*, (37), конструкции типа «*нет, (но)…*», (38):

(36) Ща́с начнёшь говори́ть «на́до отвеча́ть за всё самому́! не сва́ливать вину́ на това́рища / сосе́дей / жену́»… Ну что́? я́ всё зна́ю / я́ не ге́ний / э́то вы́ше мои́х челове́ческих си́л. Я́ / коро́ль / про́сто КОРО́ЛЬ / обыкнове́нный…

(37) где́ бы́л нарисо́ван чёрный квадра́т без вся́ких ра́мок/ про́сто зама́занный КВАДРА́Т

(38) пожа́р на спи́чечной фа́брике в Балаба́ново. — Отли́чно! Же́ртвы е́сть? — М-м… Не́т / пока́ ПРО́СТО ГОРИ́Т.

Отсылка к шкале релевантности проявляется и в упоминавшихся выше употреблениях *просто* в качестве дискурсивного коннектора: в (33)–(35), предъявляемая причина квалифицируется как менее существенная, более конкретная и более очевидная, чем та, что потенциально рассматривалась.

Замечательным образом, скалярный характер *просто*, проявляется в том, что актуально реализованное в речи значение параметра не обязательно сдвинуто в сторону минимума шкалы релевантности по сравнению с потенциально допустимым («не гений, а всего лишь король», «горит, но не опасно, без жертв» и т. д.). Это значение может сдвигаться и в сторону максимума шкалы, ср. следующие примеры:

(39) Вы́йти за́муж по́сле дву́х дне́й знако́мства / про́сто ВЕ́РХ легкомы́слия

(40) О́н оч хорошо́ ста́л игра́ть / Андре́й. Ты́ зна́ешь / така́я те́хника / про́сто НЕВЕРОЯ́ТНО.

(41) Опя́ть кто́-нибудь придёт к ва́м к у́жину / и вы́ мне́ не дади́те спа́ть со свои́ми танцу́льками. — Ма́мочка! С тобо́й про́сто с УМА́ мо́жно сойти́!

Более того, употребления *просто*, которые отсылают к максимуму параметра, как в (39)–(41) оказываются чрезвычайно частотными: в обследованной нами выборке видеофрагментов таких случаев — 27 из 100.

Итак, мы попытались показать, что (i) для *так* маркирование нерелевантности является периферийной функцией, в качестве маркера нерелевантности *так* прототипически является акцентоносителем и формирует рему самостоятельно или со спецификатором; и (ii) для *просто* маркирование нерелевантности является центральной функцией, в качестве маркера нерелевантности *просто* прототипически безударно и входит в состав рематической составляющей. Кроме того, *просто* является скалярным маркером, функция которого — соотнести актуально реализованное в речи значение некоторого параметра с потенциальным минимумом или с потенциальным максимумом на шкале релевантности. В следующем разделе мы рассмотрим сочетание *просто так* и покажем, что в качестве маркера нерелевантности оно ведет себя неаддитивно, наследуя часть свойств *просто* и часто свойств *так*.

4. Маркер нерелевантности *просто так*

По аналогии с *так* и *просто* было проанализировано и размечено расположение фразовых акцентов в ста видеофрагментах в случайной выборке из имеющихся в корпусе вхождений (201) по запросу «*просто* на расстоянии 1 от *так*». В выборке восемь фрагментов из ста содержат «побочные» случаи последовательности «*просто* + *так*», где *просто* и *так*, оказываются случайными линейными соседями:

(42) И говори́те вы́ та́к ПРО́СТО / та́к И́СКРЕННЕ

Остальные 92 вхождения демонстрируют устойчивое использование *просто так* для обозначения нерелевантности. *Просто так* наследует безударность *просто* и ударность *так*: прототипически ударным в последовательности является *так*. Исключения, как и в случае с *просто*, связаны с контрастом и/или эмфазой, что может приводить к следующим просодическим эффектам –– к смещению акцента на *просто* (8 случаев), ср. (43), (44), или к двухакцентной конфигурации типа «шляпа» — на *просто* реализуется подъем тона, а на *так* — падение, ср. (45):

(43) А я́ соску́чился без хоро́ших шофёро́в. Я́ ж не ПРО́СТО та́к. Я́ ЗАРАБО́ТАТЬ да́м.

(44) Но́вый го́д ПРО́СТО та́к не прихо́дит / его́ на́до ВСТРЕЧА́ТЬ.

(45) А быва́ет та́к / Ариста́рх Па́лыч / что́бы мо́жно бы́ло влюби́ться ПРО́СТО ТА́К / с пе́рвого взгля́да / а́?

Кроме того, имеется пять случаев, где оба компонента последовательности оказываются безударными, а эмфатический акцент реализуется на другом

слове той же коммуникативно-просодической составляющей. Например, в (46) первое вхождение *просто так* просодически реализуется двухакценто («подъем-падение»), а второе и третье вхождение являются «цитационными» и произносятся безакценто в аллегровом темпе — в контексте эмфатического акцента на *что* и *самолёт*, соответственно:

(46) Вот чё они́ э́тот самолёт угна́ли? — Да про́ст та́к. — что́ про́сто та́к? Про́сто та́к самолёт / что́ ль / угна́ли?

В остальных 75 случаях из 92 *просто так* систематически имеет акцент на *так*.

С точки зрения синтаксиса сочетание *просто так*, в отличие от *просто*, последовательно ведет себя не как частица, а как наречие/предикатив. Примерно в половине случаев (48 из 92) *просто так* формирует рематическую составляющую, имеющую адвербиальный статус:

(47) А чё он ло́х что́ ли / на́м свою́ пу́шку про́сто та́к отда́ть?

(48) А вот понима́ю то́ка не́которые слова́. Я́ люблю́ про́сто та́к слу́шать.

В 37 случаях из 92 *просто так* формирует отдельную предикацию либо изолированно, либо в сопровождении минимального набора спецификаторов:

(49) А и́м чё / все́м интере́сно ка́к у меня́ дела́? — не́т / не интере́сно. — А чё тода́ спра́шивают? — Про́сто та́к. Зде́сь вообще́ всё про́сто та́к / кроме де́нег.

В оставшихся 7 случаях из 92 *просто так* формирует парцеллированную (просодически независимую, но синтаксически встроенную в структуру предшествующей клаузы) адвербиальную группу. Статус *просто так* в этих случаях можно квалифицировать как промежуточный между отдельной предикацией и адвербиальной составляющей:

(50) я́ предлага́ю… вы́пить… по разми́ночному рюмашу́. Про́сто та́к, дру́г за дру́га.

С точки зрения семантики (не) релевантности сочетание *просто так* не наследует скалярный характер *просто*, а, подобно *так*, скорее, просто сигнализирует о том, что подразумевавшееся релевантное положение дел не имеет места. При *просто так* релевантное положение дел может непосредственно называться в тексте, например, в составе конструкций альтернативного выбора:

(51) Ну́ / мо́жет / э́то о́н про́сто та́к / а мо́жет…Мо́жет / зна́к подаёт / что́б с ни́м на свя́зь выходи́ли

(52) Скажи́те / а с Андре́ем э́то у ва́с что́ / серьёзно или про́сто та́к / а́?

Релевантный признак может вводиться и в просодически независимой адвербиальной группе, которая, однако, выступает как дублер уже имеющейся в структуре предшествующей клаузы адвербиальной группы, это — так называемая «эхо-конструкция» в терминах Кибрик, Подлесская (2009):

(53) Но никто́ ни ра́зу не позво́лил себе́ про́сто ТА́К / из ле́ни / написа́ть всё от балды́

(54) не о́чень при́нято толкова́ть апокалипси́ческие те́ксты про́сто ТА́К / профа́нно

Однако, как и в случаях с *так* и *просто*, релевантный признак может проясняться в более широком контексте. Например, *просто так* часто возникает там, где обсуждается наличие/отсутствие основания для возникновения некоторого положения дел:

(55) Ску́чно без во́дки. — А что́ / обяза́тельно напива́ться / как свинья́? — А чё ещё де́лать? — Да мо́жно про́сто ТА́К посиде́ть / поговори́ть по душа́м

(56) И вообще́ / у на́с про́сто ТА́К не сажа́ют / по́нял?

(57) Челове́ка уда́ришь? — Могу́. — За что́? — Да́же про́с ТА́К могу́.

Как видим, для *просто так* обозначение нерелевантности является основной функцией, акцентоносителем в последовательности систематически является *так*; со структурной точки зрения, *просто так* формирует отдельную предикацию или адвербиальную группу.

5. Заключение

Итак, лексические средства выражения (не)релевантности *так*, *просто* и *просто так* при очевидном сходстве семантики и дистрибуции обладают, тем не менее, заметными различиями. Прежде всего, соответствующие лексемы (в случае *просто так* — сочетание лексем) многократно различаются по представленности в корпусе: на 4 107 056 слов МУРКО приходится 29 279 вхождений *так*, 4768 вхождений *просто* и 201 вхождение *просто так*. При этом для *так* вокабулы, непосредственно отвечающие за обозначение (не)релевантности, составляют лишь около 10 %, тогда как для *просто* и *просто так* обозначение (не)релевантности — это основная функция, проявляющаяся в 92–95 % всех употреблений. С просодической точки зрения *так* в функции маркера (не)релевантности всегда является акцентоносителем, в *просто так* акцентоносителем прототипически является *так*, а *просто* прототипически безакцентно (редкие исключения возможны в контексте контраста и эмфазы). С семантической точки зрения *так* и *просто так* сигнализируют о том, что подразумевавшееся релевантное положение дел не имеет места, а *просто* имеет дополнительно

скалярную семантику, соотнося актуально реализованное в речи значение некоторого параметра с потенциальным минимумом или с потенциальным максимумом на шкале релевантности. С синтаксической точки зрения *просто* является частицей, а *так* и *просто так* функционируют как наречие/предикатив.

Разумеется, данное пилотное исследование позволило лишь в самых общих чертах охарактеризовать поведение этих единиц. Заметим, что мультимодальный корпус открывает для решения такого рода задач весьма нетривиальные возможности. Так, по нашим предварительным наблюдениям оказывается, что маркеры нерелевантности часто сопровождаются специфическими жестами. Приведем два примера. В (58) после *просто так* следует иллюстративный жест «легкомысленное порхающее дирижирование правой кистью» — в значении «не важно что»:

(58) Я́ не де́лаю про́сто ТА́К во́т / я́ де́лаю то́ / что́ мне́ НРА́ВИТСЯ / то́ / что́ ХОЧУ́

Сходным образом, но не после, а перед *просто так*, говорящий делает резкую отмашку (от себя левое предплечье сверху вниз), сопровождая этот жест отрицательным кивком — в значении «а, пусть его!»:

(59) вот та́к ща́с почу́вствовал / взя́л бы и ПОДАРИ́Л. Про́сто ТА́К.

В перспективе, мы надеемся провести более глубокое мультимодальное исследование маркеров нерелевантности, что позволит выявить в этой семантической зоне корреляцию между лексико-грамматической структурой, коммуникативно-просодической организацией речи и жестовым поведением говорящего.

Литература

1. *Булыгина Т. В., Шмелев А. Д.* (1988) Несколько замечаний о словах типа *несколько* // Язык: система и функционирование. Москва: Наука, 44–54.
2. *Булыгина Т. В., Шмелев А. Д.* (1997) Языковая концептуализация мира (на материале русской грамматики). Москва: Школа “Языки русской культуры”.
3. *Зализняк А. А.* (1980) Грамматический словарь русского языка: Словоизменение. Москва: Русский язык.
4. *Кибрик А. А., Подлесская В. И.* (Ред.). (2009). Рассказы о сновидениях: Корпусное исследование устного русского дискурса. Москва: Языки славянских культур.
5. *Падучева Е. В.* (1997) *Давно* и *долго* // Логический анализ языка. Язык и время. Москва: Индрик, 253–266.
6. *Подлесская В. И. Кибрик А. А.* (2009) Дискурсивные маркеры в структуре устного рассказа: опыт корпусного исследования // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 27–31 мая 2009 г.). Вып. 8 (15). — Москва: РГГУ, 390–395.

7. *Подлесская В. И.* (2013) Нечеткая номинация в русской разговорной речи: опыт корпусного исследования // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 29 мая — 2 июня 2013 г.). Вып. 12 (19) — Москва: РГГУ, 2013, 561–573.
8. *Янко Т. Е.* (2008) Интонационные стратегии русской речи в сопоставительном аспекте. Москва: Языки славянских культур.
9. *Янко Т. Е.* (2015) Использование корпусного материала при анализе коммуникативных функций лексем *редко* и *мало* // Вопросы филологии (в печати)

References

1. *Bulygina T. V., Shmelev A. D.* (1988) Neskol'ko zamechanij o slovax tipa *neskol'ko* [Some remarks about words of the *neskol'ko* type] // Jazyk: sistema i funkcionirovanie [Language: system and function]. Moscow: Nauka, 44–54.
2. *Bulygina T. V., Shmelev A. D.* (1997) Jazykovaja konceptualizacija mira (na materiale russkoj grammatiki) [How language conceptualizes the world (evidence from Russian grammar)]. Moscow: Shkola "Jazyki Russkoj Kul'tury".
3. *Zaliznjak A. A.* (1980) Grammaticheskij slovar' russkogo jazyka: Slovoizmenenie [A dictionary of Russian grammar: Inflection]. Moscow: Russkij Jazyk.
4. *Kibrik A. A., Podlesskaya V. I.* [Eds.] (2009) Rasskazy o snovidenijax: korpusnoe issledovanie usntogo russkogo diskursa [Night Dream Stories: A corpus study of spoken Russian discourse]. Moscow: Jazyki Slavjanskix Kul'tur.
5. *Paducheva E. V.* (1997) *Davno* i *dolgo* [*Long ago* and *for a long time*]// Logicheskij analiz jazyka. Jazyk i vremja [Logic analysis of language. Language and time]. Moscow: Indrik, 253–266.
6. *Podlesskaya V. I., Kibrik A. A.* (2009) Diskursivnye markery v structure ustnogo rasskaza: opyt korpusnogo issledovanija [The role of discourse markers in local discourse structure: a corpus study] // Computational Linguistics and Intellectual Technologies: papers from the Annual International Conference "Dialog" (Bekasovo, May 37–31, 2009), vol. 8(15). — Moscow: RGGU, 390–395.
7. *Podlesskaya V. I.* (2013) Nechëtkaja nomonacija v russkoj razgovornoj rechi [Vague reference in Russian: Evidence from spoken corpora]// Computational Linguistics and Intellectual Technologies: papers from the Annual International Conference "Dialog" (Bekasovo, May 29 — June 2, 2013), vol. 12(19). — Moscow: RGGU, 561–573.
8. *Janko T. E.* (2008) Intonacionnye strategii russkoj rechi v tipologicheskom aspekte [Intonational strategies in spoken Russian from a comparative perspective]. Moskva: Jazyki Slavjanskix Kul'tur.
9. *Janko T. E.* (2015) Ispol'zovanie korpusnogo materiala pri analize kommunikativnyx funkcij leksem *redko* i *malo* [Communicative functions of the words redko i malo: evidence from corpus data] // Voprosy filologii (in print)

LEARNING BY ANALOGY IN A HYBRID ONTOLOGICAL NETWORK

Ponomarev S. V. (serv@newmail.ru)

Sputnik LLC, Moscow, Russia

This article describes the general principles of question-answering (QA) system, which produces answers to questions by analogy with the answers and the questions at training sets. As a knowledge base the system uses a number of ontological information of words and expressions from open-access sources and statistic information, collected by processing large text corpora.

The knowledge base is presented as a hybrid ontological network—an oriented graph, where vertices[1] are the words and expressions and edges are the links between words. In addition, each link between two words or expressions is oriented, typified and weighted. The link type characterizes the information source, from which this link and its type were extracted (for example, synonym from Wiktionary). Link weight is determined by reliable information source. All links, obtained from dictionaries and ontological bases, have the weight equals to one. The links, collected by processing text corpora, have the weight equals to frequency of relevant agreed bigrams (for example, a bigram adjective + noun).

The structure of the hybrid ontological network characterizes by a large number of links between the network vertices. Besides direct links connecting two particular network vertices, there could be used composite links, passes through intermediate vertices, which leads to cardinally increasing of number of possible ways between vertices.

Here's a training algorithm that allows setting in the hybrid ontological network the links between words and items in term of combinations of weighted paths between network vertices.

Key words: ontology, linked data, query answering, semantics

1. The review of ontology systems with Natural Language interface

Systems with Natural Language interface can be divided into two groups—the first, Natural Language dialog with the user oriented (QA systems) and the second—those using Natural Language information sources to detach from the text entities and relations, for mapping into ontological databases.

QA systems, for example, QASIO [2] ontology-based domain-specific NLQA [3] and cross ontology QA on semantic Web [4], use translation of the Natural Language request into the requests for ontologies format. Both SPARQL and ones' own query languages can be used for execution of requests for ontologies.

---

[1] https://en.wikipedia.org/wiki/Vertex_(graph_theory)

The translation is realised using the formal rules that pose corresponding request template for each possible type of requests, for example:

Who	asking what or which person or people (subject)	PERSON
How far	asking about distance	NUMBER
How many	asking about quantity (countable)	NUMBER
How much	asking about quantity (uncountable)	METRICS

Thus, QA systems described realise the function of increasing of friendliness of ontology access for user, without changing the ontology data itself.

From the other side, the industry's present-day task is to translate inner technical documentation into the machine-processable form and integrate different contractor's documentation into the unified ontology, for unification of the information access.

Data Engineering Methodology of the ISO 19526 [5] standard regulates the forms of integration and processing of technical information from different sources.

The works were done of automatic parsing of Russian documents with excretion of entities and the relations between them with the use of ABBYY Compreno [6] technology. The parser accepts at the input Natural Language technical text and brings its mapping to existing ontologies.

The impossibility to eliminate all the ambiguities inherent to Natural Language, is the factor that limits possibilities of this approach. Consequently, mapping of technical text into ontologies cannot be univocal, it should represent statistically probabilistic structure.

Accordingly, methods processing such ontological data should consider the ambiguity and, perhaps, the in coordination of mapping that was built.

This paper shows the approaches that allow to substitute hand production of Natural Language request analysis rules by methods of teaching by examples. And also—approaches of solving the ontology data ambiguities by means of combination of different sources ontological data and use of relations that have probabilistic nature.

2. The Hybrid Ontological Network

Open-access ontologies in Russian do not involve the processing of links with probabilistic nature, because in these ontologies indication of the triplet weight doesn't provide. Accordingly, these ontologies can't be expanded by the links, accumulated in statistical text corpora processing, and merging them with other information sources is difficult of discrepancies between different information sources. Adding to the triplet value of its confidence (weight) allows solving these problems.

We call an ontology hybrid if:

1. It is composed of several independent sources;
2. It contains triplets (links), accumulated in statistical text corpora processing;
3. Each triplet characterized by type and weight.

The main properties of this ontology are redundancy and high relatedness. Redundancy arises from the duplication of most ontological links in various used sources, and high relatedness arises in inclusion of links obtained by statistical text processing.

If we represent this ontology as a network with vertices-concepts and edges-links, then the hybrid ontological network characterizes by a large number of possible paths between network vertices, including through some intermediate vertices. The use of statistical data ensures that not even listed concepts in used information sources, for example, rare words or names connect with other network vertices by sufficient number of links.

The total number of vertices in the network 1,355,135 and summary of link types of the hybrid ontological network are shown in Table 1.

Table 1. Structure of the hybrid ontological network

N	Link	Number of links	Link type	Source
3	Idiomatic expressions	9,334	Ontological	Wiktionary [7]
4	Epithets	49,929		
5	Antonyms	24,900		
6	Synonyms	739,053		
7	Hypernyms	29,545		
8	Hyponyms	30,871		
9	Higher category	12,332		
0	Set phrases	16,068		
13	Related words	407,895		
14	Holonymy	475		
15	Meronymy	667		
10	Categories	226,800		
24	Examples of use	16,463		
2	Defining words	4,672,480	Statistical	Defining words are words from articles of dictionaries for which the ratio of the frequency of words in the article to the frequency of words in the whole corpus ratio is as large as possible.
12	Homonym relations	17,092	Statistical	Homonym relations are set between vertices by comparing all possible word grammatical forms.
11	Words are included in one phrase	231,416,665	Statistical	Uncoordinated N-grams obtained by parsing news corpus
30	Word is adjacent to the left	22,551,832		

N	Link	Number of links	Link type	Source
17	N-gram noun + noun "house of cards"	28,722,993	Statistical	Collecting statistics using the SDK Grammatical dictionary [8]
18	N-gram adverb + verb "work hard"	2,148,646		
19	N-gram adverb + adjective "very good"	1,722,124		
20	N-gram preposition + noun "at the table"	623,370		
21	N-gram verb + managed object "see a mouse"	4,234,149		
22	N-gram adjective + noun "spiral galaxy"	10,249,513		
23	N-gram noun + verb	7,518,027		
25	The phrase is composed of	951,895	Internal	Network vertices of several words (collocations and phrases) have links with the words of which they consist.
26	The first word of the phrase	310,817		
27	The second word of the phrase	310,817		
28	Number of instance of a word to the phrase	934,023		
29	Number of instance of a collocation to the phrase	228,386		

Described structure of the hybrid ontological network allows to set the relationship between network vertices of various types, for example—"synonym" or "attribute value". At the same time, the links are characterized by computable confidence level in the range [0, 1]. In other words, established relationship is essentially a classifier that estimates whether there is a relationship between the real-world entities on the basis of available information on the network.

3. Automatic relation building

Automatic relation building is based on training sets. Relations, collected from training results, allow QA system to form the response in a manner similar to the method of forming the answer to the question in a learning sample. The right relation between question and answer in the learning sample is unknown, as there is only pair "question and answer" available for training without comments of what conclusions

have led person to this particular answer. Thus, a well-formed relation should outwardly repeat structure of human conclusions.

Relations are set paths between network vertices, encoded as an index sequence of link type. For example, the link "TOMATO >> COLOR" can be coded as follows: "TOMATO >> coherent n-grams "noun + adj."(−22) >> RED >> hypernym (7) >> COLOR". In case of arbitrary start and target vertices: «START >> coherent n-grams "noun + adj." (−22) >> RESULT >> hypernym (7) >> FINISH». At that, this network path is not the only one, and path variety between the start and target vertices may be obtained by passing through another link types. If we use composite paths passing through intermediate vertices, then the total number of possible paths between two network vertices increases like an avalanche.

Each of the paths may be weighted by appropriate coefficient. Thus, the path leading to the correct result may have an increased coefficient, as paths that do not lead to a correct result—have a reduced coefficient or being deleted.

Let's look at example. Given a triple values "TOMATO", "RED", "COLOR".

Table 2. The link structure between the vertices: "TOMATO", "RED", "COLOR"

	Number of links	**Links lead to "RED"**
TOMATO	23	1. Related n-grams "adjective + noun", back link (−22).
COLOR	30	1. Hyponyms (8); 2. Related n-grams "adjective + noun", back link (−22); 3. Phrase consists of, back link (−25).

Link № 8 "hyponyms" obtained by Wiktionary parsing, which explicitly set the connection "Red is a hyponym of the word Color". Link № 22 is derived by statistical text processing, as agreed bigram "red" and "color". As we move from the word "color" to the word "red", the back link is used (from the color to red). Link № 25 shows that the word "color" and "red" are present together in one of the network vertices. Vertice "Red color" received by parsing dictionaries.

Thus, if we follow from the vertice "TOMATO" to link № −22 and from the vertice "COLOR" to links № 8, № −22 and № −25, then we'll get the following link picture:

Table 3. Building a path between the vertices: "TOMATO", "RED", "COLOR"

	Link type	**Number of links**	**Top 10 of the vertices**
TOMATO	−22	382	**red**, best, fresh, ripe, rotten, sliced, marinated, rotten, green, salty ...
COLOR	8	7	blue, purple, sea color, orange, **red**, brown, green
COLOR	−22	4,032	whole, **red**, white, black, yellow, green, blue, gray, such a, own ...
COLOR	−25	265	versicolour, zinnwaldite, fanal, Black Sea, surah, old gold, cream, blue dust, dark tangerine, light-color ...

To build a path in the hybrid network, you should specify a set of pairs "link type—link weight":

START >> (–22, 1) >> RESULT;
FINISH >> (8, 1/3) >> RESULT;
FINISH >> (–22, 1/3) >> RESULT;
FINISH >> (–25, 1/3) >> RESULT.

Apply this path to different sets of arguments.

Table 4. Applying constructed path in the network

START	FINISH	RESULT
TOMATO	COLOR	red—0.3744 green—0.2380 most—0.1914 blue—0.1709 fresh—0.1691
CURRANT	COLOR	black—1.1153 red—1.0118 green—0.1941 blue—0.1670 brown—0.1555
CAR	COLOR	red—0.1915 own—0.1789 green—0.1719 blue—0.1683 brown—0.1556
SEA	COLOR	black—0.2317 red—0.2293 blue—0.2233 green—0.1836 mediterranean—0.1570
SEA	SIZE	length—0.2732 high—0.2732 width—0.2500 depth—0.2500 black—0.2026

As Table 4 shows the generated path gives satisfactory results for requests related to the color of the object, but does not apply to other types of requests such as the request of size. For the path formed by the only learning triple " TOMATO >> RED >> COLOR" it is natural. Try to expand the rule to train it also on the triple "SEA >> LARGE >> SIZE". In general, it could not pick up such weights for pairs "link type—link weight" without passing through intermediate vertices that outgoing rule satisfactorily completes work on requests associated with both requests: the color and the size.

Let's consider the building of paths, passing through one intermediate vertice.

Table 5. The link structure through one intermediate vertice

	All link types	**Links lead to "RED"**
TOMATO	606	60
COLOR	1,018	177
	All link types	**Links lead to "LARGE"**
SEA	1,016	107
SIZE	807	108

Further increase the number of intermediate vertices leads to an avalanche-like increase of available links. A lot of links increase the chances of learning algorithm to generate an effective way to respond to a wide class of requests.

Consider the algorithm of the path construction on the hybrid network:

1. Given: one or more training triples "START >> RESULT >> FINISH";
2. Build links such as "START >> RESULT" and "FINISH >> RESULT";
3. Select such weight rates of links, that the desired value "RESULT" was maximum;
4. Carry out a test run: in a path specific values "START" and "FINISH" are substituted of training triples and verify that the maximum value of "RESULT" is the value of teaching triple. If the condition is satisfied, then the path is ready and we exit from the algorithm.
5. Build relations "START >> RESULT" and "FINISH >> RESULT" through one additional vertice;
6. Repeat from step 3.

4. QA system

QA system is trained on pairs "question—answer" given in Russian. Firstly we produce syntactic analysis of question and build syntactic tree. Each type of syntactic trees corresponds one rule at the rule base in QA system. If there is no rule found under questions with the same syntactic tree, then the rule is formed. If the corresponding rule is found, then firstly response is generated. After, the answer is compared with the correct answer, and if they do not match, the rule is extended by another pair of "question-answer" and it's being relearned.

The rule contains the information, which words in question we use as arguments "START" and "FINISH", and information which word in the correct answer use as "RESULT". The bespoke correct answer is broken into words, and for each word is determined by its type: 1) the function words (pronouns, verbs, punctuation), 2) the transfer word (present as in the question and in the answer) 3) the computable word (there isn't in the question, but it can be derived from the question words by building relationships). Thus the system leaves the function words in its place for generation the answer, replaces the transfer words on the relevant words from the question, calculates the computable words and then aligns with the grammatical phrase attributes

(gender, number, case). In case of disambiguate, what words to use as arguments to "START" and "FINISH", calculations are carried out for all the variants, and then the system select the path with the least number of links.

Let's look at the example of this approach. We form the rule basis; each rule encodes the output method of answer to a question by analogy with examples from the training set. When you start the system rule base is empty. Each question of the system corresponds the correct answer. The system tries to generate their own answer according to rules base, in case of failure—it remembers a new "question—answer" pair. At the same time, memorized pairs of "question—answer" create a new rule of inference or specify an existing one.

Step 1: The rule base is empty, so random response generated. A pair of "question—answer" memorized.

(1) *Question: Какой глубины лужа?* (What depth is the puddle?)
Correct Answer: Лужа—мелкая. (The puddle is small)
Generated Answer: Глубина. (Depth)
New Rule Added.

Step 2. In the rule base there's the only rule obtained in step 1, and the system tries to apply this rule to the question. Attempt fails and the rule is corrected.

(2) *Question: Какой глубины море?* (What depth is the sea?)
Correct Answer: Море—глубокое. (The sea is deep)
Generated Answer: Море—мелкое. (The sea is small)
Adding 1 New Path.

Step 3. In the rule base there's still the only rule, but it's taught at two examples. The system makes a successful attempt to apply this rule to the question. Thus, in this case two training examples are enough to obtain practically valuable rule.

(3) *Question: Какой глубины океан?* (What depth is the ocean?)
Correct Answer: Океан—глубокий. (The ocean is deep)
Generated Answer: Океан—глубокий. (The ocean is deep)
Correct Answer Found.

Step 4. The syntactic structure of pair "question—answer" is changed, so the use of the existing rule does not give the correct result. Another rule is generated.

(4) *Question: Какой глубины лужа?* (What depth is the puddle?)
Correct Answer: Лужа маленькой глубины. (The puddle is small depth)
Generated Answer: Лужа—мелкая. (The puddle is small)
Generated Answer: Глубина. (Depth)
New Rule Added.

Step 5. Attempt to apply rule № 2, obtained in step 4, gives the correct result within meaning, but not coinciding exactly with the correct answer. Rule № 2 is corrected.

(5) *Question: Какой глубины море?* (What depth is the sea?)
Correct Answer: Море большой глубины. (The sea is deep depth)
Generated Answer: Море огромной глубины. (The sea is vast depth)
Adding 1 New Path.

Step 6. The syntactic structure of pair "question—answer" corresponds more with the rule № 2, than with the rule № 1. The attempt to apply rule № 2 to determine the color instead of the size gives the expected result.

(6) *Question: Какого цвета огурец?* (What color is the cucumber?)
Correct Answer: Огурец зеленого цвета. (The cucumber is green color)
Generated Answer: Огурец зеленого цвета. (The cucumber is green color)
Correct Answer Found.

Similar way the appliance of rule № 2 gives the correct answers to the questions "What color is a tomato?" and "What size is a seed?". The structure of rule № 2 rules is given in Table 6. Total number of paths in rule № 2 is 54 left and 123 right, the most important paths are included to Table 6.

Table 6. The structure of rule №2

START	Weight of path	Path	RESULT	Weight of path	Path	FINISH
color	0.2352160	26 0 7	green	0.1685550	3 −7 27 9	cucumber
depth	0.1176080	25 0 7	red	0.1348440	12 −9 24 −16	tomato
size	0.1176080	28 0 7	big	0.1348440	−15 −9 24 −16	seed
	0.0996208	7 −16	small	0.0374567	3 −3 26 24	sea
	0.0958917	3		0.0345754	5 −27 −32 24	puddle
	0.0740494	3 27 0 −8		0.0345754	−8 −27 −32 24	
	0.0282505	−25 27 0 −8		0.0313329	−25 15 7	
	0.0270013	6 0 −10 −23		0.0280925	2 −10 −27 5	
	0.0270013	−9 0 −10 −23		0.0232490	4 −7 27 9	
	0.0270013	−28 0 −10 −23		0.0210694	3 27 −29 12	

The data represented in Table 6 is interpreted as follows: for getting the word "green" from the word "color" we need to build the path "COLOR >> the first word of the phrase (26) >> Set phrases(0) >> hypernyms (7) >> GREEN". Or we can use shorter path "COLOR >> idiomatic expressions (3) >> GREEN". Not all the paths formed the rule № 2 can be built for each pair of arguments "color + cucumber", "depth + puddle" and etc., but the excess amount of paths guarantees to find a sufficient number of paths to separate the correct result. Negative indexes mean back links, so link № 7 is a link from hypernym to hyponym. This link is different from link № 8, because the used data source (Wiktionary) is not complete, and an essential part of back links is not filled.

Let's take a detailed look at the first three paths of links "color—green," "depth—big" and "size—small". As seen, the first link type 25, 26 and 28 is an internal link type between the phrases and their components, words (see Table 1). The basic phrases related to the words "color", "depth" and "size" are listed in Table 7.

Table 7. Some links of the hybrid network

	Color	**Depth**	**Size**
The phrase is composed of (25)	white; painting; flowering; blue; number; yellow; protective; green;	container; seriousness; abyss; solidity; significance; thoroughness; serious; depth	height; growth; coverage; border; length; volume; scale; measure;
The first word of the phrase (26)	hair color; skin color; color of languages; aquamarine-colored; turquoise-colored;	depth on languages; depths of the earth; depth of hold; depth of inhale; nesting depth;	size on languages; yield; size of the female pelvic organs in the sagittal section;
Number of instance of a word to the phrase (28)	verbs discoloration; verbs color development; blue color; yellow color; yellow color;	languages; deep; deeply; deep; in ancient days;	measure; size adverbs; size on languages; measure by language; enormous size;

For the word "color" links 25 and 28 give the required "green", but the 26th link does not lead to an acceptable result. On the other hand, for the words "depth" and "size" can be seen accordance only with the 28th link type: "at a depth," "large size", and the 25th and 26th links do not lead to direct result. It demonstrates the network redundancy and the rules formed as a set of paths in the network. That means fixity rules in their application to the arguments that have not all affixed link types. On the other hand, the rules in the paths passing through the 26th link means that even the naked eye cannot see sense, a positive effect on the productivity of the final rule turns out.

It is important that the ontological data sources, which the ontological network formed, do not contain links such as “sea—deep” and “cucumber—green.” These links obtained by statistical text processing methods.

Technology demonstrator of deduction by analogy is available at http://servponomarev.livejournal.com/6059.html

5. Quality control of QA system

Rule № 2 (Table 6) as a result of learning in two examples “What depth is the puddle?” and “What depth is the sea?” used to assess the response quality to a set of questions oriented at getting the typical response of an attribute object value. Rule № 2, trained only on “depth”, is used to demonstrate the possibility of generalizing to other types of attributes.

Table 8. The answers to some questions according to rule №2

What taste is the lemon?	The lemon is tart.
What taste is the watermelon?	The watermelon is sweet.
What taste is the herring?	The herring is pungent.
What taste is the onion?	The onion is strong.
What weight is the grain?	The grain is small.
What weight is the cobble?	The cobble is small.
What weight is the bar-bell?	The bar-bell is small.
What color is the cucumber?	The cucumber is green.
What color is the strawberry?	The strawberry is bright.
What color is the lemon?	The lemon is bright.
What color is grime?	Grime is deep.

As seen from Table 8, rule №2, trained by attribute “depth” satisfactorily fulfils also the attribute “taste”, and in some cases the attribute “color”. However, to obtain high-quality results, we should set rules individually for each of the attribute types. For example, the rule formed by the pair question-answer “What color is snow? Snow is white.” shows the following results in mode of relearning according to correct answers.

Table 9. The answers to some questions according to rule “color”

What color is the cucumber?	The cucumber is green.
What color is grime?	Grime is black.
What color is the cloud?	The cloud is black.
What color is the cloud?	The cloud is gray.
What color is the sky?	The sky is grey.
What color is the grass?	The grass is green.
What color is the tomato?	The tomato is green.
What color is the lemon?	The lemon is green.

Performance of large-scale testing hindered by the lack of context, which allows to select the one concrete correct value from the list of valid values. So the answer that the lemon is green is allowable, although more common answer is "The lemon is yellow." In future versions of QA system we will plan introduction context recording.

6. Quality control of automatic relation building

The method of automatic relation building described in paragraph 2 used in "The First International Workshop on Russian Semantic Similarity Evaluation" [1], where in the category "Evaluation based on Semantic Relation Classification" was obtained accuracy in 0.9209 on criterion Area under Curve (AUC), which ensured 3rd place in the competition.

7. Follow-up research

The research efforts in the direction of automatic selection of the syntactic form of answer to question using only the statistics dialogs without learning by example. We plan to create QA system that generates answers to questions, taking into account the context in its natural form, like a dialogue between two people.

References

1. *Panchenko A., Loukachevitch N. V., Ustalov D., Paperno D., Meyer C. M., Konstantinova N.* (2015) "RUSSE: The First Workshop on Russian Semantic Similarity". In Proceeding of the Dialogue 2015 conference. Moscow, Russia
2. *Moussa, A. M., Abdel-Kader, R. F.* "QUASIO: A Question Answering System for YAGO Ontology". International Journal of Database Theory and Application 4(2), 99–112 (2011)
3. *Athira P. M., Sreeja M. and P. C. Reghuraj* "Architecture of an Ontology-Based Domain Specific Natural Language Question Answering System". International Journal of Web & Semantic Technology (IJWesT) Vol. 4, No. 4, October 2013
4. *Lopez, V., Uren, V. S., Sabou, M., Motta, E.* "Cross Ontology Query Answering on the semantic Web: an Initial Evaluation." In: Gil, Y., Noy, N. F. (eds.) K-CAP, pp. 17–24 ACM (2009)
5. ISO 15926 Reference Data Engineering Methodology, http://techinvestlab.ru/files/RefDataEngenEnglish/RefDataEngen_ver_3_English.doc
6. The reference data extraction from technical texts in natural languages, http://www.slideshare.net/vvagr/reference-dataextraction
7. Wiktionary, https://ru.wiktionary.org/
8. Russian and English Morphology for Windows and Linux, http://solarix.ru/grammatical-dictionary-api-en.shtml

ПОИСК И РАНЖИРОВАНИЕ ИЛЛЮСТРИРУЮЩИХ ПРИМЕРОВ ДЛЯ ПЕРЕВОДНОГО СЛОВАРЯ

Протопопова Е. (rhubarb@yandex-team.ru),
Антонова А. (antonova@yandex-team.ru),
Мисюрев А. (misyurev@yandex-team.ru)

Яндекс, Москва, Россия

Ключевые слова: автоматическое создание словарей, параллельный конкорданс, векторные модели

ACQUIRING RELEVANT CONTEXT EXAMPLES FOR A TRANSLATION DICTIONARY

Protopopova E. (rhubarb@yandex-team.ru),
Antonova A. (antonova@yandex-team.ru),
Misyurev A. (misyurev@yandex-team.ru)

Yandex, Moscow, Russia

This paper addresses the problem of automatic acquisition of parallel context examples for a translation dictionary. We extract them automatically from a parallel corpus, relying on word alignments and parse trees. The ranking of the extracted examples is an essential problem, since we need to select the most distinctive and informative contexts. We propose a machine learning approach as an alternative to simple ranking criteria, such as frequency, or mutual information. We perform the analysis of common sources of inadequate context examples and design a set of features, which can possibly distinguish the bad examples from the good ones. We also experiment with vector models (word2vec) in order to get features that are sensitive to semantics. The evaluation result show that the best of our ranking methods yields 31% improvement in accuracy compared to the ranking by frequency, and 20% improvement over the ranking by mutual information. Using vector models also improves the classification performance.

Keywords: bilingual dictionary extraction, bilingual concordance, vector models

1. Introduction

The paper is concerned with a problem of automatically acquiring the illustrative translation examples for English-Russian machine dictionary. Such examples can enrich the dictionary entry, illustrate semantic and syntactic selectional preferences, and help the user to differentiate between the meanings of multiple translation variants. Many well-known translation dictionaries include examples, which had been prepared by professional lexicographers.

Recently, the growing amount of parallel documents in the Internet and the current progress in language processing algorithms makes it possible to retrieve the context examples automatically from large-scale parallel corpora. Figure 1 shows how automatically extracted context examples are used to illustrate different meanings of the words *'French'* and *'пример' ('example')* in an online dictionary.

Fig. 1. Illustrative examples in a bilingual machine dictionary

The dictionary format imposes the following requirements on the context examples:

- Only one or several best examples are shown per one translation.
- Examples should be short well-formed grammatical phrases.
- Examples should represent a characteristic use of a given word or expression.

Examples are extracted from parallel sentences where a given translation pair is found with the help of word alignment (acquired by GIZA++ [9]) and a phrase extraction algorithm [6]. The sentences are processed by a dependency parser [1] so that we can search for words in different forms. Only phrases constituting a connected subgraph of a sentence parse tree are extracted and thus most of the ungrammatical phrases are discarded. This step is discussed in [2]. Parallel corpus is compiled from web-archives of a commercial search engine.

The essential problem is the ranking, since we need to eliminate all kinds of noisy contexts and select the most distinctive and informative ones. There exist simple ranking criteria, such as frequency, or mutual information, but they do not always work well. For example, when a phrase frequency is taken into account, then frequent but

useless examples are often ranked best (*then <go> → затем <перейти>*). If we use a metric like mutual information, too specific examples can be scored better (*unpregnant <woman> → небеременная <женщина>*).

In this paper we propose a machine learning approach to the ranking problem. We analyse typical mistakes and design a set of features, which can possibly distinguish the bad examples from the good ones. We also use features from vector models (word2vec tool [8]) in order to predict syntactic and semantic relatedness between words.

We report on the experiments with two-words examples (bigrams). Different classifiers are trained on a manually annotated sample of automatically extracted examples. The classifiers' scores are used for elimination of noisy examples and for ranking the remaining ones. In some experiments the remaining examples are ranked according to a simple measure such as frequency. We also try to estimate prediction confidence using a combination of classifiers in order to find the most relevant examples.

The results are evaluated as follows. We compute precision, recall and accuracy of the classification using an annotated test set. We also perform a comparative evaluation of the accuracy of one-best examples found by different methods. The best of our ranking methods yields 31% improvement in accuracy compared to the ranking by frequency, and 20% improvement over the ranking by mutual information.

The advantages of automatic approach to the task of creating context examples are the following:

- The automatic approach enables us to find up-to-date and frequently used phrases.
- The procedure can be repeated on bigger or different corpora in order to cover more meanings and words.
- Our statistical approach can be applied to any language pair with available corpora and a syntactic parser.

The paper is organized as follows. In Section 2 we briefly outline the related work. Section 3 describes the principles and the results of the examples annotation. Then we discuss the classification task in Section 4. Section 5 is devoted to classification experiments and system evaluation.

2. Related work

The papers concerning bilingual lexicon acquisition pay little attention to the problem mentioned in this paper, but task in general corresponds to that of building a bilingual concordance, i.e. finding all the examples of the word usage in text with their respective translations. Such systems are intended for translators and language learners. In some papers ([5], [7]) the issue is reduced to finding all sentences with a given source word and the presented systems do not take into account target expression and do not extract smaller phrases.

Ranking is not of great importance when building a bilingual concordance. Some of the systems such as the one discussed in [4] provide user with frequency information about collocations. In [10] the system ranks sentences and their translations according to frequency statistics, while the authors of [3] use Dice coefficient to show more relevant translations first.

3. Examples annotation

The classification task requires annotated data for learning, so first of all the data for annotation should be prepared. In this section we describe our experimental set as well as the principles of annotation.

3.1. Selecting translation pairs and examples

In order to make training and test sets more representative we try to select translation pairs and the respective examples so that their frequency distribution reflects the real word frequency distribution in parallel corpora. It is also important to illustrate source words which are more frequently queried in machine dictionary. We have noticed [2] that the amount of queries for source words highly correlates with source word frequencies, so we can rely on corpus statistics when selecting pairs for annotation. Finally, we create a random sample of English words excluding the most frequent hundred.

Each source word has one or several translations (target expressions) in our dictionary. For each pair 'source word—target expression' we extract all possible context examples from a web-based parallel corpus. However, random sampling from all examples would be quite unreliable because it would not ensure the balance between relevant and irrelevant examples. Thus, for each translation pair we select several best examples according to source and target frequencies as follows:

$$F = log(f_3) - log(f_1) - log(f_2)$$

where f_1 and f_2 are frequencies of words which do not form a given translation pair and f_3 is the whole example frequency.

3.2. Annotation principles

Table 1. Annotation principles

score	both sides annotation	one-side annotation
1	both parts are meaningless and grammatically incorrect; the parts are not translation equivalents	a phrase is meaningless and grammatically incorrect
2	one of the parts can be scored with 1 in one-side annotation or one or both parts are grammatically incorrect	a phrase is grammatically incorrect; a phrase is not a translation equivalent
3	both parts are grammatically correct but do not reflect any peculiarities of the translation pair	a phrase is grammatically correct but does not reflect any peculiarity of a word/expression
4	both parts are correct and partially illustrate peculiarities of a given pair	a phrase is correct and partially illustrates peculiarities of a given word/expression
5	relevant example	relevant example

The machine dictionary is created automatically and contains some noise. These noisy translation pairs and the respective examples are removed from the annotation set. Then we perform two kinds of annotation: assessing the whole example and assessing its source and target phrases separately. In each case we assign a score which ranges from 1 (very bad) to 5 (excellent). Table 1 specifies the requirements for all scores. The examples scored with 3 are then removed from the training set, as they are neither negative, nor positive.

3.3. Annotation results

After annotating 700 bigram examples we remove phrases extracted for incorrect translation equivalents. The number of examples for each score is shown on Figure 2. The number of erroneous Russian examples is somewhat higher because of the higher number of grammatical mistakes (see Section 3.4). As a whole, more positive examples were extracted due to filtering by frequency.

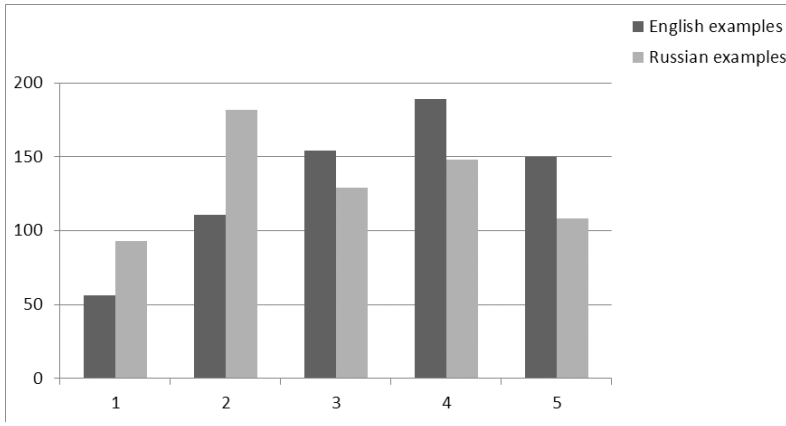


Fig. 2. The distribution of scores

3.4. Error analysis

The following errors are observed in automatically extracted examples (source and target expressions are marked with angle brackets, errors are marked with an asterisk):

1. Inadequacy in surface form
   (a) Ungrammatical phrases
   **<preparation> enamel → <составление> эмали*
   *<appreciate> acrobatics → *<оценить> акробатика*
   (b) Incomplete phrases
   county <detention> → деревенский <исправительный>

(c) Phrases not in dictionary form
**<created> tsunamis → *<породило> цунами*
*monstrously <big> → *чудовищно <огромная>*
*header files → *заголовочных файлов*

(d) Phrases containing a foreign word
*<improve> resiliency → *<улучшать> resiliency*
*unformatted <capacity> → *unformatted <емкость>*
**<beginning> shvatyvanija → начало> схватывания*

(e) Phrases containing a misspelled word
*caribbean <community> → *караибское <содружество>*
*burgundy <sole> → *бардовая <подошва>*

2. Inadequacy in meaning

(a) Uninformative phrases
**его <любовь> → *his <fondness>*
**очень <глупый> → *really <stupid>*
**nonpregnant <woman >→ *небеременная <женщина>*

(b) Phrases with unrelated words
**pickled <loveliness> → *маринованная <красота>*
**<saving> neurotic → *<спасение> невротиков*
**синхроничная <жизнь> → *synchronistic <life>*

(c) Hardly understandable phrases with specific meaning
**sagittal <reconstruction> → *сагиттальная <реконструкция>*
**threshold <panel> → *пороговое <табло>*

(d) Machine translation
**<soya> squirrels → <соевый> белок*
**<character> stitches → <символьные> строчки*
**harvest <control> → жмешь <контрол>*
**Berners-<whether> → Бернерс-<ли>*
*hi <camcorder> → *привет <видеокамеры>*

(e) Offensive contexts for neutral words
naked <girl> → голая <девушка>
<Japanese> militarists → <японские> милитаристы
Hitlerite <Germany> → гитлеровская <Германия>
<become> a Shaheed → <стать> шахидом

(f) Phrases which are not translations of each other
<saving> rolling → <спасение> утопающих

The first group of errors can be explained by the fact that almost no limitations are placed on extracted parse subtrees. This problem may be overcome by means of special rules which filter out some ungrammatical translations. Parallel machine translated sentences and misspelled words are frequent on websites and can be drawn when gathering parallel corpus in the internet. In some cases the sentences in the target text contain only partial translation of the source sentences, and phrases from them are also extracted as context examples.

4. Classification

4.1. Feature sets

We propose several groups of features which can distinguish the irrelevant examples from the informative ones.

Language model scores (LM)

Language models are concerned with example fluency as well as with filtering out grammatically incorrect expressions. We use English and Russian trigram language models compiled on big monolingual corpora containing Web documents. We also build part-of-speech trigram models using the sequences of morphological tags acquired by a statistical parser. We compute the following values:

- example perplexity according to unigram LM (2 features);
- example perplexity according to trigram (bigram in case of bigram examples) LM (2 features);
- the scores mentioned above using part-of-speech LM (2 features).

Relative Frequency (RelF)

We use the example frequency as described in Section 3:

$$RelF = log(f_3) - log(f_1) - log(f_2)$$

where f_1 and f_2 are frequencies of words which do not form a given translation pair and f_3 is an example frequency.

Mutual information (MI)

The average mutual information score for bigrams is computed for both sides of example treating two words as bigram if there is a syntactic link between them:

$$MI = log \frac{f(w_1, w_2)}{f(w_1) f(w_2)}$$

where $f(w)$ is the relative frequency of word w in a corpus and $f(w_1, w_2)$ is the relative frequency of the pair (w_1, w_2) connected with an arc in a parse tree. The relative frequencies are extracted from monolingual corpora annotated with the help of a statistical parser. Thus we can find more idiomatic examples with less frequent words.

Semantic similarity (Sim)

Word vectors computed by word2vec tool [8] on a large monolingual corpus have proved to be very efficient in capturing different linguistic regularities. We try to exploit them to find out more typical and specific word usages. Using word2vec tool we represent each word by a 200-dimensional word vector. Then we compute the cosine similarity measure in each one-side example. In case of three or more words we suggest calculating average similarity between all vectors as well as similarity between a given word and all other words in an example. This results in two features, one for each example side.

Vector models (WV)

As mentioned above, each word can be represented as a semantic vector, which can be used for training as is. We concatenate all vectors for words in a one-side example and also introduce binary features to indicate a key word for an example. Thus a feature vector for a two-word example (u,v) where the key is the second word looks like $(u_1, ..., u_{200}, v_1, ..., v_{200}, 0, 1)$ which means that 402 features are used. Concatenation requires that examples of different length are trained separately.

4.2. Classifiers

Simple binary classification

The examples annotation is quite detailed and quite difficult to predict automatically, so first of all we build a binary classifier to distinguish between informative and irrelevant or erroneous examples. We use a Random forest classifier as well as a feed-forward neural network with a single hidden layer.

Estimating prediction confidence

The multilabel annotation is useful when we try to find the examples which are undoubtedly relevant. For this purpose we combine four binary random forest classifiers for each score excluding examples with the closest score from the training set, for instance, when treating the 4th class as positive examples, we remove all examples of the 5th class and leave 1st and 2nd classes as negative examples. When predicting scores on test set, we use all classifiers and choose that with the highest predicted value and estimate confidence c as

$$c = \left| max(f_1, f_2) - max(f_4, f_5) \right|$$

where f_i is a predicted value of i-th classifier.

5. Test data and experimental setup

5.1. Assessing classifiers

Table 2. Classifiers performance. P_0 is the precision on negative examples and A is the classification accuracy

	$P_0(en)$	$A(en)$	$P_0(ru)$	$A(ru)$
RF_1	0.71	**0.74**	0.62	**0.64**
RF_2	**0.83**	0.65	**0.63**	0.62
NN	0.67	0.65	0.56	0.61

One-side prediction

	P_0	A
RF_3	0.690	0.70
RF_4	0.685	0.71

Both sides prediction

For each of the 52 random English words sampled according to the frequency distribution paired with all possible Russian translations from an online machine dictionary [2] we extract 3 best examples according to both sides frequencies and annotate the resulting examples removing those for incorrect translations. We split the resulting set into training (416 examples) and test (206 examples) parts. Firstly, we perform classification for source and target side separately using the following combinations:

- RF_1—random forest classifier using *WV* features;
- RF_2—combination of four random forest classifiers using the same feature set;
- *NN*—neural network using the same feature set.

Classifiers performance is shown in table 2a. We compute precision measure on negative examples to check whether our method is useful in eliminating erroneous and irrelevant contexts. We can notice that the results on English sides of examples are slightly better. This may be explained by the quality of word vectors which should be trained on larger corpus for languages with rich inflection.

Secondly, we use features for both sides to classify full examples. We apply random forest classification to the following feature sets:

- RF_3—*LM, MI, RelF* and *Sim* features;
- RF_4—all the features described in section 4.1.

Table 2b shows the evaluation results. The learning curves for RF_1 and RF_3 are presented on Figure 3.

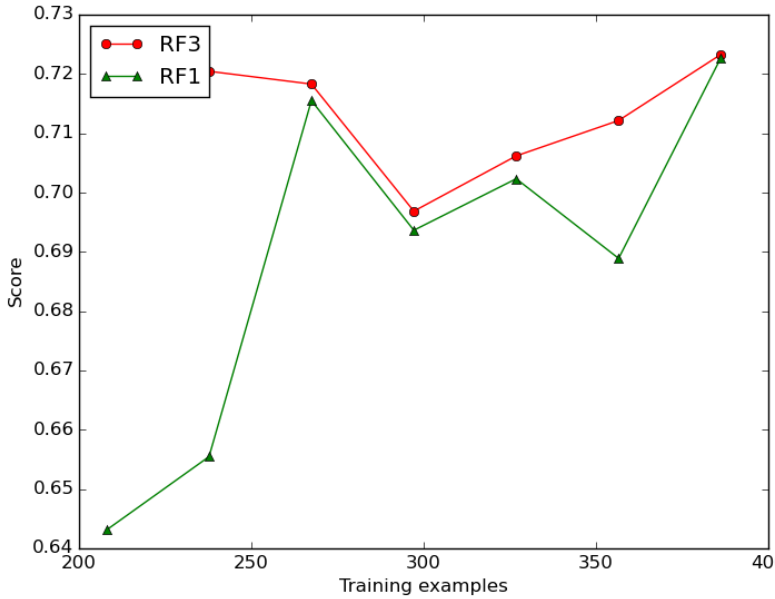


Fig. 3. Accuracy score for training sets of different size

5.2. Comparison with existing methods

Table 3. Number of correct examples extracted from different rankings

	correct examples	**percentage of correct examples**
MI	60	42.8
F	44	31.4
RF_1	59	42.1
RF_2	76	54.3
RF_3	**88**	**62.9**
RF_3	74	52.9

For comparative evaluation we choose 140 translation pairs, which were not annotated for the training set and extracted all possible context examples and selected top ones according to absolute example frequency *F* (i.e. f_1 in *RelF* formula from section 4.1) and *MI* metric described in Section 4.1. We compute MI for English and Russian phrase separately and then rank examples with respect to sum of scores for both sides.

We apply the same classification schemes to the resulting 22,375 examples and select the most relevant according to the following ranking:

- After RF_1, RF_3 and RF_4 classification we rank examples according to their scores (from 0 to 1).
- As mentioned before, the results of RF_2 classification include confidence scores for all values. When examples marked as good are found, we rank them according to their confidence score. When good examples appear only in one language, we select the pair with a positive value (4,5) and the highest confidence in one language and negative value (1,2) with the lowest confidence in another.

The results are shown in table 3. It can be observed that applying machine learning results in a noticeable improvement in examples quality. Examples acquired by RF_3 and selected according to the frequency ranking are compared in Table 4.

Table 4. Resulting examples, selected according to RF_3 and *F* scores

Key pair	RF_3	***F***
size—формат	standard size—стандартный формат	different sizes—различных форматов
control—контролирование	control costs—контролирование расходов	obstacle control—контролирование препятствий
guy—мужчина	white guy—белый мужчина	burly guy—дородный мужчина

Comparing different feature sets we can see that the most successful one is the one used by RF_3 classifier. These results in general correspond to those presented in Table 1 and Table 2, except for RF_2 classifier, which was expected to provide better results.

Taking into account feature importances computed by random forest we find out that the most important group is the *Sim* group. The direct comparison between word vectors (cosine similarity) seems to be the most relevant criterion, performing better than the internal vector comparison (when we use *WV* features). Using *WV* features with other groups proves to be redundant, although they would probably perform better on a larger training set.

The proposed confidence score improves the classification accuracy as compared with simple regression (RF_2 vs. RF_1). It would be interesting to apply this approach to classifiers with other groups of features.

6. Conclusion

We have described the procedure for automatically acquiring relevant illustrative translation examples for English-Russian machine dictionary. We have analyzed errors in phrases extracted from a parallel corpus in order to find out what features should be taken into account when choosing proper examples for a bilingual dictionary and discussed the drawbacks of straightforward approaches to ranking context examples. We have described our machine learning approach to detecting the most informative examples.

We have presented the results of classification and ranking evaluation. The comparison with simple methods proves that our approach overcomes such ranking functions as frequency or mutual information and may be successfully used for examples extraction. Some of the features proposed require minimal linguistic software so that the approach may be applied to other language pairs.

References

1. *Antonova, A., Misyurev, A.* (2012). Russian dependency parser SyntAutom at the DIALOGUE-2012 parser evaluation task. In Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog 2012” [Komp’yuternaya Lingvistika i Intellektual’nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii “Dialog 2012”].
2. *Antonova, A., Misyurev, A.* (2014). Automatic Creation of Human-Oriented Translation Dictionaries. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog 2014” [Komp’yuternaya Lingvistika I Intellektual’nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii “Dialog 2014”].
3. *Bai, M.-H., Hsieh, Y.-M., Chen, K.-J., Chang, J.* (2012). DOMCAT: A Bilingual Concordancer for Domain-Specific Computer Assisted Translation. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics. Jeju, Republic of Corea.

4. *Barlow, M.* (2004). Parallel Concordancing and Translation. Translating and the Computer.
5. *Kjaersgaard, P. S.* (1987). RefTex—a context-based translation aid. In D., Copenhagen University of Copenhagen (Ed.), Third conference of the European Chapter of the Association for Computational Linguistics: Proceedings of the conference.
6. *Koehn, P.* (2007). Moses: Open Source Toolkit for Statistical Machine Translation.
7. *Langlois, L.* (1996). Bilingual concordancers: a new tool for bilingual lexicographers. In Expanding MT horizons: Proceedings of the Second Conference of the Association for Machine Translation in the Americas. Montreal, Quebec, Canada.
8. *Mikolov, T., Chen, K., Corrado, G., Dean, J.* (2013). Efficient Estimation of Word Representations in Vector Space. In Proceedings of Workshop at ICLR.
9. *Och, F. J., Ney, H.* (2003). A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics, 29(1), 19–51.
10. *Wu, J.-C., Yeh, K. C., Chuang, T. C., Shei, W.-C., Chang, J. S.* (2003). TotalRecall: A Bilingual Concordance for Computer Assisted Translation and Language Learning. In ACL-2003: 41st Annual meeting of the Association for Computational Linguistics. Sapporo, Japan.

ИЗВЛЕЧЕНИЕ ИНФОРМАЦИИ ИЗ КЛИНИЧЕСКИХ ТЕКСТОВ НА РУССКОМ ЯЗЫКЕ

Шелманов А. О. (shelmanov@isa.ru),
Смирнов И. В. (ivs@isa.ru)
Институт системного анализа Российской академии наук, Москва, Россия

Вишнёва Е. А. (vishneva@nczd.ru)
Научный центр здоровья детей, Москва, Россия

Ключевые слова: обработка клинических текстов, извлечение информации, клинические тексты, извлечение атрибутов заболеваний, медицинские тексты

INFORMATION EXTRACTION FROM CLINICAL TEXTS IN RUSSIAN

Shelmanov A. O. (shelmanov@isa.ru),
Smirnov I. V. (ivs@isa.ru)
Institute for Systems Analysis of Russian Academy of Sciences, Moscow, Russia

Vishneva E. A. (vishneva@nczd.ru)
Scientific Centre of Children Health, Moscow, Russia

We present and evaluate the pipeline for processing of clinical notes in Russian. The paper addresses the tasks of drug identification and disease template filling, which are related to entity recognition and relation extraction. The disease template filling consists in recognition of disease mentions in text, mapping them to concepts of a thesaurus, and discovering their attributes. Discovering attributes means identifying corresponding spans in text, linking them to diseases, and normalizing them i.e. determining their generalized meaning from a predefined set. We implemented tools for determining the following attributes of disease mentions: negation; the flag indicating the disease mention is not related to a patient; severity; course; and body site. For different tasks, we used different techniques: rule-based patterns and several supervised machine-learning methods. Since there were no annotated corpora of clinical notes in the Russian language available for research purposes, we annotated a dataset, which we used for training and evaluation of the developed tools. The created corpus is available for researchers through the data use agreement.

Keywords: clinical text processing, information extraction, annotated corpus, clinical narrative, disease template filling, medical text, EHR

1. Introduction

A vast amount of clinical data is stored as free text. Electronic health records of medical facilities accumulate radiology, echocardiography, and electrocardiogram reports, anamnesis, results of ultrasound diagnostics, discharge summaries, and many other types of notes related to patient healthcare history that are written in a natural language. This is a very rich knowledge source, which is still difficult to exploit because of its unstructured nature. Reworking it into computable form can benefit the biomedical research, patient medical history management, and eventually improve the healthcare. Although there are many natural language processing techniques developed for information and knowledge extraction, the specificity of clinical narrative and tasks arising in the medical domain facilitate the development of specialized methods, language resources, and tools. It is a promising and fruitful scientific direction in natural language processing.

Much of the research in this direction is focused on processing English clinical texts. However, it is also important to create tools and resources for other languages. In the current research, we present and evaluate the pipeline for processing of clinical notes in Russian. The paper addresses the tasks of drug identification and disease template filling, which are related to entity recognition and relation extraction. The disease template filling consists in recognition of disease mentions in text, mapping them to concepts of a thesaurus, and discovering their attributes. Discovering attributes means identifying corresponding spans in text, linking them to diseases, and normalizing them i.e. determining their generalized meaning from a predefined set. We implemented tools for determining the following attributes of diseases: negation; the flag indicating the disease mention is not related to a patient; severity; course; and body site. For different tasks, we used different techniques: rule-based patterns and several supervised machine-learning methods. Since there were no annotated corpora of clinical notes in the Russian language available for research purposes, we annotated a dataset, which we used for the training and evaluation of the developed tools.

2. Related work

Natural language processing of medical texts is a rapidly developing research area. A big number of challenges conducted in the last few years that are devoted to the problems of clinical and biomedical information retrieval and text processing reflects the growing interest in this area for academic community. We briefly review some of them related to information extraction from English clinical texts. 2012 i2b2 /VA Challenge was focused on temporal relation extraction from clinical narratives (Sun et al., 2013). Pilot ShARe/CLEF eHealth 2013 Evaluation Lab set tasks of identifying in clinical reports disease mentions and acronym abbreviations as well as mapping them to thesaurus (Suominen et al., 2013). On ShARe/CLEF eHealth 2014 Evaluation Lab these tasks were extended to disease template filling—identification of disease mentions and their attributes (severity, course, body site etc.). SemEval 2014 Task 7 was similar to CLEF eHealth 2013 and was aimed at disease and acronym/abbreviation identification and

normalization[1]. Sem Eval 2015 Task 6 was about temporal information extraction from clinical texts[2]. SemEval-2015 Task 14 was similar to CLEF eHealth 2014 disease template filling task[3]. We should also note CLEF eHealth 2015 challenge that although is not directly related to clinical text processing, but shows that efforts are not limited to medical information extraction for English language. It introduces the task of named entity recognition in French biomedical texts[4]. This initiative became possible because of recent creation of French annotated biomedical corpus Quaero (Névéol et al., 2014).

The problems of processing clinical notes in Russian we focus on in the current work are largely similar to the tasks of entity recognition and disease template filling introduced by the CLEF eHealth 2014 challenge. The majority of the CLEF eHealth participants applied to this task both the rule-based approaches and supervised machine-learning techniques for extracting different attributes, e.g. (Hamon et al., 2014; Liu and Ku, 2014; Huynh and Ho, 2014). (Ramanan and Nathan, 2014) applied to this task purely rule-based methods. In (Mkrtchyan and Sonntag, 2014), approach based on deep dependency parsing was implemented. Several participants adapted and expanded the cTAKES (Savova et al., 2010) medical text-processing framework (Sequeira et al., 2014; Johri et al., 2014). The participants commonly exploited the MetaMap (Aronson and Lang, 2010) and NegEx (Chapman et al., 2013) tools for mapping found terms to thesaurus concepts and negation detection correspondingly.

Besides, the problem of disease template filling was recently addressed in (Dligach et al., 2014). The severity and body site attributes were linked to disease annotations with SVM classifier. The researchers tested several kernels including the novel tree-based kernel, and found that the common radial-based kernel was suitable for the aforementioned tasks. The evaluation against rule-based baselines showed the advantage of machine learning techniques.

Since there was no basic toolchain for processing clinical texts in Russian, we had to create it from the scratch. Our medical term (diseases, drugs, symptoms, body sites) identification and normalization module is an analogue of MetaMap. The module for disease negation detection and the module for detection that a disease is not related to a patient apply approaches implemented in NegEx. Although our modules are based on somewhat similar approaches to those implemented in the tools for English text processing, we made some modifications in them that take into account peculiarities of the Russian language. In modules for discovering severity and course attributes of disease mentions as well as for linking body sites to disease mentions, we implemented the state-of-the-art machine learning techniques. The annotated corpus we are currently developing contains Russian clinical free-text notes annotated with treatment, symptom, drug, body sites, disorder / disease mentions, their attributes and relations. The closest analogous for English are the corpus of the

---

1 http://alt.qcri.org/semeval2014/task7/

2 http://alt.qcri.org/semeval2015/task6/

3 http://alt.qcri.org/semeval2015/task14/

4 https://sites.google.com/site/clefehealth2015/

Shared Annotated Resources (ShARe) initiative (Pradhan et al., 2015) and the corpus of Strategic Health IT Advanced Research Project: Area 4 (SHARPn)[5].

3. Methods for processing of clinical texts in Russian

The developed clinical text-processing pipeline begins with basic NLP analysis. For tokenization, sentence splitting, part-of-speech tagging, and lemmatization we used well-known pipeline for Russian from AOT.ru (Sokirko, 2001). The dependency syntactic parsing was performed by MaltParser (Nivre et al., 2007) trained on SynTagRus (Apresjan et al., 2005) with configuration described in (Nivre et al., 2008; Sharoff and Nivre, 2011; Smirnov et al., 2014). We find the performance of both of these tools on clinical texts satisfactory. The next step in the pipeline is medical term identification and normalization. It is followed by the disease/symptom negation detection and determining whether disease mentions are related to a patient or not. We will refer to the latter task as "not patient" flag detection. The pipeline concludes by discovering severity and course attributes of disease mentions and linking body sites to disease mentions.

3.1. Identifying medical terms in text and mapping them to thesauri

To solve many tasks of clinical and biomedical text processing it is necessary to perform the basic identification of medical terms in text, term normalization, and mapping them to semantic types. For these tasks, we developed a heuristic- and thesaurus-based module. We used two thesauri: UMLS Metathesaurus (Schuyler et al., 1993) for disease, symptom, and body site identification; a thesaurus based on State Register of Drugs (SRD)[6] for the drug identification.

UMLS Metathesaurus is an extensive compendium of medical lexicons, classifications, code sets, and thesauri. The main feature of it is that the concepts with the similar sense from different knowledge sources are mapped to the single CUI (concept unique identifier) in the metathesaurus. UMLS also provides Semantic Network that maps concepts into one or more coarse-grained semantic types (McCray, 1989). The only Russian thesaurus present in UMLS is MeSHRUS[7].

SRD is a database of all drugs officially registered and allowed for sale in Russia. Records in SRD among other information contain trade names of drugs and their active chemicals. Since many drugs have different trade names but almost the same compound and can substitute each other in medication, we preprocessed the database and grouped drugs with similar active chemicals into concepts of thesaurus.

The method implemented in our module was mainly inspired by MetaMap—a widely used linguistically motivated tool for mapping terms from medical texts to concepts

---

[5] http://informatics.mayo.edu/sharp/index.php/Main_Page

[6] http://grls.rosminzdrav.ru/

[7] http://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/MSHRUS/

in UMLS Metathesaurus. This tool can be applied only to English text because it strictly relies on handcrafted heuristics and English lexis. The developed module for processing texts in Russian is also rule-based. It generates extensive amount of term variants from text expressions, performs fuzzy comparison between the variants and the thesaurus terms, then ranks the variants by heuristically reasoned score, and picks the most confident ones.

In the first step, the parser scans the text for single keywords from the thesaurus using its reverted index, which is preliminarily built by mapping lemmatized tokens of thesaurus terms to thesaurus concepts. Stop-words like particles, punctuation, prepositions are pruned from the index. The found keywords become the seed term variants for the second step—generation of complex variants.

The generation procedure constructs new term variants from a keyword by expanding it with nodes of the syntactic tree and with words from linear context around the keyword. The procedure is driven by heuristics and constrained by parser parameters (maximum syntax tree depth, maximum number of nodes above the keyword, window of the linear context etc.).

The third step is the assessment of how the generated variants correspond to the thesaurus terms. For each token in the variant, the parser determines the sets of thesaurus terms that correspond to the token and unite them in a single set. Then parser assesses the similarity between the variant and terms from the set. The assessment is a value between 0 (the weakest match) and 1 (the strongest match), which is computed as a linear convolution of three components: “lexical involvement”, “centrality”, and “cohesiveness” (values between 0 and 1). Although the sense of these components is somewhat similar to analogues implemented in the MetaMap, we compute them in the different way.

The “lexical involvement” assesses how the tokens of the variant correspond to the tokens of the thesaurus term. The component ignores order of tokens and considers terms as a bag of words. The lexical involvement is a weighted harmonic mean F_β of two values: the total weight of tokens of a variant among tokens of a term from the thesaurus and the total weight of tokens of a term from the thesaurus among tokens of a variant. Unlike the MetaMap, which convolves these values linearly, we used harmonic mean to combine them because it more strictly penalizes the component if any of them has a small value.

The “centrality” shows whether the most significant token of a variant from syntax perspective (the head of the phrase) is present in the thesaurus term. It is 1 if the syntactic head of the biggest phrase in the variant is present in the term of the thesaurus and 0 otherwise. This component can prune variants with many less significant token matches that lack the most important match of the syntactic head.

The “cohesiveness” extends the idea of “lexical involvement” to syntactically connected phrases. It is a weighted harmonic mean of two values “coherence” and “variant coverage”. The “coherence” is a maximum ratio of a number of syntactically connected tokens in the variant that participate in the match with the thesaurus term and a number of tokens in the variant. This component can prune variants with matches that are not syntactically linked and therefore are not related.

Finally, variant—thesaurus term pairs are selected by a threshold and filtered by heuristics. The chosen variants are mapped to the semantic types according to their identifier with the biggest score.

3.2. Negation and "not patient" flag detection

It is crucial to know whether the found disease or symptom mentions are negated in text and whether they are associated with the patient or with another person, e.g. patient's relative, since many clinical notes also contain information about patient's heredity.

For negation detection, medical text processing systems usually apply simple approach based on pattern matching. For example, NegEx implements an algorithm that searchers for a list of patterns in a linear context in a window around a disease mention. Despite the simplicity of the algorithm, it proved itself robust with moderate performance and was adopted in cTAKES.

Besides the aforementioned strategy, we also implemented pattern search in a syntax tree, which showed somewhat better performance. The latter approach was integrated into the final pipeline. The developed negation detection module searches for the following patterns: particle "не" ("not") syntactically depends on one of the tokens of the disease/symptom; particle "не" ("does not") syntactically depends on predicate (single predicate word of complex predicate with auxiliary verb) that governs a token of the disease/symptom term; particle "нет"("no") governs a token from the disease/symptom term; a token of the disease/symptom term is governed by negation predicate, e.g. "отсутствует" ("is absent"); particle "нет" ("no") immediately follows a disease/symptom mention.

Module that detects "not patient" flag mainly searches for mentions of patient relatives and patterns that associate them with disease terms in text: "у" + "relative mention" syntactically connected to or precedes disease term in a sentence; "наследственность" ("heredity") precedes disease mention in a sentence.

3.3. Discovering severity, course, and body site attributes of diseases

Severity attributes capture and normalize text cues that clarify whether the disease mentioned in text is "slight", "moderate", or "severe". Course attributes capture and normalize text cues about disease progress: "worsened", "changed", "improved", "resolved". Body site attributes determine which body part disease mentions are associated with in a clinical text. In example: "*Диагностирована астма, атопическая, легкое персистирующее течение, ремиссия. … Имеется ангиопатия сетчатки.*" ("*Diagnosed asthma, atopic, mild persistent, during remission. … Retinal angiopathy is present.*"). The string "*легкое персистирующее течение*" ("*mild persistent*") should be captured in the severity attribute of the disease mention "*астма, атопическая*" ("*asthma, atopic*") and normalized as "slight"; the string "*ремиссия*" ("*during remission*") should be captured in the course attribute of the same disease mention and normalized as "improved"; the string "*сетчатки*" ("retinal") should be marked as body site and linked to the disease mention "*ангиопатия*" ("*angiopathy*"). The common practical usage of the extracted information is summarizing electronic health records for information or analytical systems in terms of standardized format like Clinical Document

Architecture[8]. For the tasks of discovering severity, course, and body site attributes, we implemented several modules based on supervised machine learning methods.

The modules that discover severity and course attributes are almost the same. They consist of two separate submodules for the attribute span identification with linking it to the corresponding disease mention and for the attribute normalization. For the given disease mention, the submodule for the attribute span identification scans tokens of a sentence, in which the corresponding disease mention is located, and applies to them binary classifier that predicts whether token is an attribute related to the disease mention or not. When every token is classified, the tokens marked by the same attribute are grouped into continuous annotations. We used the following lexical and syntactic features: lemmas and postags of tokens in a window around the classified token; whether the classified token syntactically depends from the disease term; distance between the classified token and the disease term; relative position of the token regarding to the given disease mention; number of disease annotations between the disease mention and the token. When attribute spans are identified, another submodule with a separate classifier normalizes them. The feature set of the latter classifier is composed of token lemmas lying in the corresponding span of severity or course annotation represented as a bag of words.

In the task of discovering body site attributes, the spans that represent body parts are identified by thesaurus-based parser described in section 3.1. Therefore, only linking these spans to disease mentions is required. The module for linking body sites to diseases scans all body site—disease pairs within a sentence and analyzes them with a binary classifier. The features of the classifier include distance in tokens between a disease mention and a body site, whether they are syntactically linked, whether they are attached to the same word (e.g., predicate), the postag of this word, the number of disease mentions between the given disease mention and the body site.

We tested several classifiers for each of the tasks and subtasks: linear SVM, SVM with radial basis kernel, random forest, and AdaBoost. The parameter tuning was performed on the developed corpus via cross-validation.

4. Annotated corpus of clinical notes in Russian

In conjunction with specialists of Scientific Center of Children Health (SCCH)[9] we created annotated corpus of clinical free-text notes in Russian. The corpus is based on medical histories of more than 60 SCCH patients with allergic and pulmonary disorders and diseases. It comprises discharge summaries, radiology, echocardiography, ultrasound diagnostics reports, recommendations, and other records created by different physicians. The documents in the corpus were de-identified: all names were removed and dates were altered. With the help of SCCH experts, we developed an annotation scheme and a guideline. The scheme encompasses span annotations: “Disease”, “Symptom”,

[8] http://www.hl7.org/implement/standards/product_brief.cfm?product_id=7

[9] http://www.nczd.ru/eng/

"Drug", "Treatment", "Body location", "Severity", "Course"; attributes: "Negation" (for "Disease", "Symptom"), "NotPatient" (for "Disease", "Symptom", "Treatment"), "Degree" (for "Severity"), "CType" (for "Course"); relations between "Severity" and "Disease", "Course" and "Disease", "Body location" and "Disease"; and other annotations.

The corpus was annotated and verified by physicians. For annotation purposes, we used Brat—the web-based tool, which was originally created for BioNLP challenge[10]. The Fig. 1 illustrates a fragment of an annotated text.

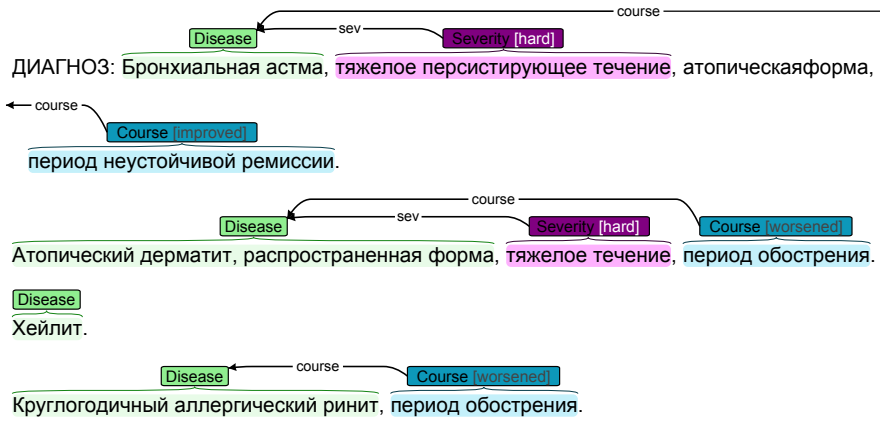


Fig. 1. Example of annotated text

The corpus currently contains more than 112 fully annotated texts with almost 45,000 tokens. There are more than 7,600 annotated entities and more than 4,000 annotated attributes and relations. The work on the corpus is still in progress: we are adding new texts and planning to expand the annotation scheme.

The corpus is freely available for the research community through the data use agreement[11]. However, the human subjects training certificate[12] would be required for access to it since the corpus contains the medical data of real patients.

5. Experiments

5.1. Disease and drug identification performance

The testing of medical term identification module was performed on a randomly selected holdout consisting of 30 texts; the rest of the corpus was used for parameter tuning. We calculated relaxed versions of precision, recall, and F_1 score. In the relaxed assessment, a span overlapping a gold standard span is considered correct.

---

[10] http://brat.nlplab.org/

[11] http://nlp.isa.ru/datasets/clinical

[12] https://phrp.nihtraining.com/users/login.php

Especially for the disease identification task, we prepared two baselines. The first baseline marks in text all words of thesaurus concepts related to "disease" semantic type. This baseline tends to maximum recall. The second baseline marks in text only token chains that exactly match a whole bag of words of a thesaurus concept. This baseline tends to maximum precision. Table 2 presents performance of the developed module and the baselines.

Table 2. Performance of the disease identification module and the baselines

Module	Recall,%	Precision,%	F_1-score,%
Disease identification	72.8	95.1	**82.4**
Baseline 1	**84.9**	9.3	16.7
Baseline 2	69.8	**99.2**	81.9

As expected, the implemented module has better overall performance than the baselines, however, it appeared to be very close to the performance of the baseline 2. The developed method should more significantly outmatch the baseline 2 in the task of mapping terms to concepts because the baseline 2 often maps terms to the very general concepts although they should be mapped to the more specific ones. For example, the module finds term "аллергический ринит" ("allergic rhinitis") in text. This string cannot be mapped exactly to the bag of words of any thesaurus term, so the baseline 2 will not find it. However, it will find "ринит" ("rhinitis") and map it to the more general concept. Both answers will be considered as a match for the disease identification task, however, in the concept-mapping task, the answer of the module will be the only correct one.

For the evaluation of the drug identification performance, we used exactly the same framework as for the disease identification. The precision was 84.3, the recall was 74.6, and the F_1-score was 79.2. The recall is somewhat lower than expected because annotators marked not only registered drugs but also mentions of therapeutic cosmetics (e.g., anti-allergic cream). We also note that some false negatives were due to corpus texts contained contractions and general names of drugs like "пенициллин" ("penicillin"), which is not present in SRD, instead of full name like "бензилпенициллин" ("benzathine penicillin"), which is in SRD.

5.2. Symptom negation and "not patient" flag detection performance

We found that the annotated corpus contains just a few negations of disease mentions, much more negations are related to symptoms. Therefore, instead of disease negation detection we only evaluated the performance of the symptom negation detection.

The number of negation and "not patient" flag annotations in corpus is relatively small—both around 100. To make the test set representative, we had to evaluate negation and "not patient" flag detection using the whole corpus rather than the selected holdout. Since the developed modules are rule-based, such approach has a minor effect on the performance assessment bias. In the evaluation, we only took into account attributes of symptom and disease mentions identified by our system to exclude false negatives of the term identification module (table 3).

Table 3. Performance of the symptom negation and "not subject" detection

Module	Recall,%	Precision,%	F_1-score,%
Symptom negation	98.7	95.3	97.0
Disease "not patient"	90.9	96.8	93.8

The obtained results show that a rather simple approach and few patterns can cover the most cases of negation and "not patient" attributes in clinical texts. However, we admit that the current test set is small and not very representative for the perfect evaluation.

5.3. Performance of discovering severity, course, and body site attributes of diseases

For all evaluations, we used 5-fold cross validation on the annotated corpus.

We separately evaluated performance of the identification and normalization of severity and course attributes. To exclude the false negatives of the disease identification module, we only took into account course and severity annotations that are related to disease mentions identified by our system. For evaluation of severity and course identification, similarly to the evaluation of medical term identification, we calculated relaxed versions of precision, recall, and F_1-score (tables 4 and 5).

Attribute normalization is a multilabel classification task without "empty" class. Therefore, evaluation of severity and course normalization was performed via accuracy (table 6).

The performance of linking body sites to diseases was evaluated via relaxed precision, recall, and F_1-score. Only disease mentions that were identified by the system were taken into account (table 7).

Table 4. Performance of severity identification

Classifier	Recall,%	Precision,%	F_1-score,%
Linear SVM	99.2	41.7	58.6
RBF SVM	95.0	80.8	87.1
Random forest	93.6	82.6	**87.5**
AdaBoost (Dec. tree)	97.3	75.2	84.7

Table 5. Performance of course identification

Classifier	Recall,%	Precision,%	F_1-score,%
Linear SVM	92.3	99.2	**95.7**
RBF SVM	88.3	99.3	93.4
Random forest	88.3	99.3	93.4
AdaBoost (Dec. tree)	90.0	98.4	93.9

Table 6. Performance of severity and course normalization

Module	Classifier	Accuracy,%
Severity normalization	Linear SVM	88.4
	RBF SVM	88.0
	Random forest	89.3
	AdaBoost (Dec. tree)	**89.8**
Course normalization	Linear SVM	89.4
	RBF SVM	91.4
	Random forest	**92.7**
	AdaBoost (Dec. tree)	91.4

Table 7. Performance of linking body sites to diseases

Classifier	Precision, %	Recall, %	F_1-score, %
Linear SVM	85.4	77.5	81.0
RBF SVM	91.4	76.6	**83.3**
Random forest	86.6	75.8	80.8
AdaBoost (Dec. tree)	84.0	76.6	79.9

Experiments with different classifiers showed statistical difference between their results only in the task of severity identification. This can be because of importance of word collocations that are differently modeled by these classifiers. The developed modules represent a solid baseline. However, there is still a big space for improvement, which can be achieved by developing richer set of features and applying new machine learning methods. We consider that the obtained results can be a useful landmark for the future research.

6. Conclusion and future work

We presented the pipeline for processing of clinical texts in Russian and the corpus of annotated clinical notes. We evaluated the pipeline and showed that it can successfully solve the key tasks of information extraction from clinical narrative. In the ongoing work, we are expanding annotated corpus to make it more representative and suitable for training machine-learning algorithms, as well as for reliable testing of clinical text processing methods and tools. In the future work, we are planning to upgrade corpus annotation scheme and include in the pipeline modules for discovering treatments with modifiers and temporal expressions related to disease mentions. We also are planning to apply the developed pipeline for the high-level task of clinical information retrieval and clinical data analysis.

Acknowledgments

We are grateful to experts of Scientific Center of Children Health for help on annotating corpus of clinical texts in Russian. This work was supported by RFBR, project 13-04-12062.

References

1. *Apresjan J. D., Boguslavskij I. M., Iomdin B. L., Iomdin L. L., Sannikov A. V., Sannikov V. G. and Sizov L. L.* (2005), Syntactically and semantically annotated corpus of Russian language: Present state and perspectives [Sintaksicheski i semanticheski annotirovannyj korpus russkogo jazyka: sovremennoe sostojanie i perspektivy], National Corpus of Russian Language: 2003–2005 [Natsional'nyj korpus russkogo jazyka: 2003–2005], pp. 193–214, (in Russian)
2. *Aronson, A. R. and Lang, F.-M.* (2010), An overview of MetaMap: historical perspective and recent advances, Journal of the American Medical Informatics Association (JAMIA), (3), Vol. 17, pp. 229–236
3. *Bos, L. and Donnelly, K.* (2006), SNOMED-CT: The advanced terminology and coding system for eHealth, Studies in health technology and informatics, Vol. 121, pp. 279–290
4. *Chapman, W. W., Hilert, D., Velupillai, S., Kvist, M., Skeppstedt, M., Chapman, B. E., Conway, M., Tharp, M., Mowery, D. L. and Deleger, L.* (2013), Extending the NegEx lexicon for multiple languages, Studies in health technology and informatics, Vol. 192, pp. 677–681
5. *Dligach, D., Bethard, S., Becker, L., Miller, T. A. and Savova, G. K.* (2014), Discovering body site and severity modifiers in clinical texts, Journal of the American Medical Informatics Association (JAMIA), pp. 448–454
6. *Hamon, T., Grouin, C. and Zweigenbaum, P.* (2014), Disease and Disorder Template Filling using Rule-based and Statistical Approaches, In CLEF (Working Notes), pp. 79–90
7. *Huynh, H. N. and Ho, S. L. V. B. Q.* (2014), ShARe/CLEFeHealth: A Hybrid Approach for Task 2, In CLEF (Working Notes), pp. 103–110
8. *Johri, N., Niwa, Y. and Chikka, V. R.* (2014), Optimizing Apache cTAKES for Disease/Disorder Template Filling: Team HITACHI in the ShARe/CLEF 2014 eHealth Evaluation Lab, In CLEF (Working Notes), pp. 111–123
9. *Lipscomb, C. E.* (2000), Medical subject headings (MeSH), Bulletin of the Medical Library Association, Vol. 88, pp. 265–266
10. *Liu, Y.-C. and Ku, L.-W.* (2014), CLEFeHealth 2014 Normalization of Information Extraction Challenge using Multi-model Method, In CLEF (Working Notes), pp. 124–132
11. *McCray, A. T.* (1989), The UMLS Semantic Network. In Proceedings of the Annual Symposium on Computer Application in Medical Care, pp. 503–507, American Medical Informatics Association
12. *Mkrtchyan, T. and Sonntag, D.* (2014), Deep Parsing at the CLEF2014 IE Task, In CLEF (Working Notes), pp. 138–146

13. *Névéol, A., Grouin, C., Leixa, J., Rosset, S. and Zweigenbaum, P.* (2014), The Quaero French medical corpus: A resource for medical entity recognition and normalization, In Proceedings of LREC BioTxtM 2014 Fourth Workshop on Building and Evaluating Resources for Health and Biomedical Text Processing
14. *Nivre, J., Boguslavsky, I. M. and Iomdin, L. L.* (2008), Parsing the SynTagRus treebank of Russian, In Proceedings of the 22nd International Conference on Computational Linguistics, pp. 641–648, Association for Computational Linguistics
15. *Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryigit, G., Kübler, S., Marinov, S. and Marsi, E.* (2007), MaltParser: A language-independent system for data-driven dependency parsing, Natural Language Engineering, (2), Vol. 13, pp. 95–135
16. *Pradhan, S., Elhadad, N., South, B. R., Martinez, D., Christensen, L., Vogel, A., Suominen, H., Chapman, W. W. and Savova, G.* (2015), Evaluating the state of the art in disorder recognition and normalization of the clinical narrative, Journal of the American Medical Informatics Association (JAMIA), (1), Vol. 22, pp. 143–154
17. *Ramanan, S. V. and Nathan, P. S.* (2014), Cocoa: Extending a Rule-based System to Tag Disease Attributes in Clinical Records, In CLEF (Working Notes), pp. 150–155
18. *Savova, G. K., Masanz, J. J., Ogren, P. V., Zheng, J., Sohn, S., Kipper-Schuler, K. C. and Chute, C. G.* (2010), Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): Architecture, component evaluation and applications, Journal of the American Medical Informatics Association (JAMIA), (5), Vol. 17, pp. 507–513
19. *Schuyler, P. L., Hole, W. T., Tuttle, M. S. and Sherertz, D. D.* (1993), The UMLS Metathesaurus: Representing different views of biomedical concepts, Bulletin of the Medical Library Association, (2), Vol. 81, 217–222
20. *Sequeira, J., Miranda, N., Goncalves, T. and Quaresma, P.* (2014), TeamUEvora at CLEF eHealth 2014 Task2a, In CLEF (Working Notes), pp. 156–166
21. *Sharoff S. and Nivre J.* (2011), The proper place of men and machines in language technology: Processing Russian without any linguistic knowledge, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2011", pp. 591–604
22. *Smirnov, I. V., Shelmanov, A. O., Kuznetsova, E. S. and Hramoin, I. V.* (2014), Semantic-syntactic analysis of natural languages. Part II. Method for semantic-syntactic analysis of texts [Semantiko-sintaksicheskij analiz estestvennyh jazykov Chast' II. Metod semantiko-sintaksicheskogo analiza tekstov], Artificial Intelligence and Decision Making [Iskusstvennyj intellekt i prinjatie reshenij], (1), pp. 95–108, (in Russian)
23. *Sokirko, A.* (2001), A short description of Dialing Project, available at: http://www.aot.ru/
24. *Sun, W., Rumshisky, A. and Uzuner, O.* (2013), Evaluating temporal relations in clinical text: 2012 i2b2 challenge, Journal of the American Medical Informatics Association (JAMIA), (5), Vol. 20, pp. 806–813
25. *Suominen, H., Salanterä, S., Velupillai, S., Chapman, W. W., Savova, G., Elhadad, N., Pradhan, S., South, B. R., Mowery, D. L., Jones, G. J. et al.* (2013), Overview of the ShARe/CLEF eHealth evaluation lab 2013, In Information Access Evaluation. Multilinguality, Multimodality, and Visualization, Lecture Notes in Computer Science, Vol. 8138, pp. 212–231, Springer

ON SUMMARIZATION SUPPORTING READABILITY AND TRANSLATABILITY

Sheremetyeva S. O. (lanaconsult@mail.dk)

LanA Consulting ApS, Copenhagen, Denmark
National Research South Ural State University, Chelyabinsk, Russia

The article describes a methodology of developing an interactive computer system for supporting a single document text-to-text summarization process focusing on providing for high readability and translatability of the generated summary that, in turn, facilitates further human or automatic processing of the summary text, translation being the most important. The decisions on content selection is delegated to a human but are largely supported by the system. High readability and translatability of the generated text is provided by controlling the syntax of the nascent summary. The approach is a combination of empirical and rational NLP techniques and incorporates a language independent algorithm and language-dependent knowledge base. The validity of the approach was proved by its implementation into a summarizer for scientific papers in the domain of mathematical modelling in the Russian language. The summarizer is fully operational. The methodology presented in this paper is highly portable and allows for extending the summarizer to other domains and languages.

Keywords: summarization, readability, translatability, machine processing, interactivity

1. Introduction

Development of efficient summarization systems is of special importance for scholars and developers. With the cumulative total, which is estimated to pass 50 million scientific papers [9] researchers can only keep abreast of the growing number of scientific developments through summary digest. While it is generally recognized that summaries should be created operatively and contain all the key content of a full document, such properties of a summary as readability and translatability often escape both the authors and summary systems developers' attention. Good readability means that a text is easy to understand for a human, especially for a human translator who, as a rule, does not possess enough of domain knowledge and needs to clearly understand the syntactic dependences of the original text. High translatability means that a summary can be well "understood" by a machine translation system as translators and especially the authors, more and more nowadays use MT tools. It is not uncommon to see summary texts that are very problematic both for humans and machines, such as the following:

(1) *Строятся и исследуются аналитическими методами математические модели напряженных состояний тонкостенных цилиндрических оболочек, продольных, поперечных и спиральных менее прочных слоев (прослоек) в них, в том числе содержащих дефекты, более прочных слоев с дефектами, при нагружении оболочек внутренним давлением и осевой силой.*

While the terminology problem can be solved with correct terminological lexicons, it is such linguistic phenomena as long sentences, distant dependencies, complex syntactic structures, coordination, etc., that lower the levels of texts readability and translatability [21].

In this paper we describe our effort to develop a methodology for a real life tool that could support high readability and translatability in the summarization process. The approach is a combination of empirical and rational natural language processing techniques including interactive computer-supported content elicitation and fully automatic text generation. The methodology is realized in a summarization tool for research papers on mathematical modelling in the Russian language.

2. Related work and motivation

Major summarization strategies fall into extraction, abstraction or mixed paradigms and are realized in the frame of linguistic, statistical or hybrid approaches.

In linguistic approaches text summarization is considered as a transformation process by which knowledge representation structures, as generated by a natural language processing (NLP) system, are mapped into conceptually more abstract, condensed knowledge structures that account for document content[7,11]. Knowledge representation structures such as, *Schemata*, [16], *Rich Semantic Graphs* [6] or *predicate-argument representations* [18], to name just a few, are then transformed into a summary text by natural language generation (NLG) techniques. Linguistic approaches are normally abstracting and can provide higher quality summaries, but they are knowledge–heavy and suffer coverage problem. In pure statistical, normally extracting, summarization the most relevant sentences of a document are extracted according to a certain statistical metrics, e.g., a classical tf-idf term weighting scheme, and inserted in the summary as such [1, 15, 20]. Statistical systems though providing for coverage can still produce problematic output. Most popular now are mixed-paradigm hybrid approaches that on top of statistical calculations rely on a shallow to deep knowledge bases and more or less interwoven extraction and transformation components. The latter modifies selected fragments with one or several of the following techniques,—paraphrase induction [22], fragment recycling [5], sentence simplification [24], sentence compression [4], sentence fusion [2, 3] or predicting the selection of lexical units and their position in the summary [15,17]. The knowledge base in such systems can contain from lists of predicates, to rhetorical and sentence templates [23]. Summary-relevant fragments in the input are normally spotted by building conditional models over some statistical features, for example, commonly occurring word

co-occurrences in summary sentences [14, 23] or words across adjacent sentences in particular semantic roles [2,10].

We aim at developing a real world application for authors and editors that do not only guarantee the correct content of a summary but also takes care of the summary text structure, making it highly readable and translatable. So far we were not able to find any research which would combine text compression with readability/translatability issues. However, now, when summaries are the priority types of documents that undergo further processing (e.g., human or machine translation) readability and translatability are of great importance. This is especially relevant for the Russian language, whose rich morphology and free word order allow though grammatically correct, but still low readable/translatable long, syntactically extremely complex and ambiguous sentences. The authors trying to make their summaries more informative and much compressed often abuse these features of Russian that causes a lot of problems in further processing of the documents. Instructions only, and simple style checkers that are not so far sufficiently developed, especially for Russian, do not serve the purpose. Let alone that a good style does not always guarantee high translatability. It actually can be the other way around. Analysis of research in automation of text processing tasks shows that machines are better at syntax than at semantics, and systems, in which the division of labour is "semantics for humans, syntax for the computer" allow for much better results [12,13]. We therefore make our system interactive and delegate content selection to the user, while (unlike style checkers) providing extensive linguistic support and full automation of text restructuring based on natural language generation techniques. The knowledge and processing rules in our approach as in other most promising research [8, 22, 23] draw heavily on domain restrictions that contribute a lot to the system viability.

3. Approach Overview

The overall architecture of the summarizer is shown in Figure 1. The system consists of

- Domain-tuned knowledge base
- automatic noun phrase (NP) extractor, NP and predicate phrase (VP) chunkers
- interactive content elicitation module
- automatic content representation module
- automatic summary text generator
- user-friendly interface

Human intervention occurs at the stage of content selection and is linguistically supported by (i) alerting the user about the content of an initial document with explicitly marked noun and predicate terminology (see Fig. 2) and (ii) providing lexical menus and knowledge elicitation templates. The results of knowledge elicitation are further automatically processed into an underlying representation followed by the summary generation in a syntactically controlled language providing for high readability and translatability of the summary. The approach is of hybrid nature and includes language- and domain-independent algorithms that run over language-dependent

domain-tuned lexical knowledge. This makes the methodology portable across domains and languages. The details are further described on the example of a summarizer for the domain of mathematical modelling in the Russian language.

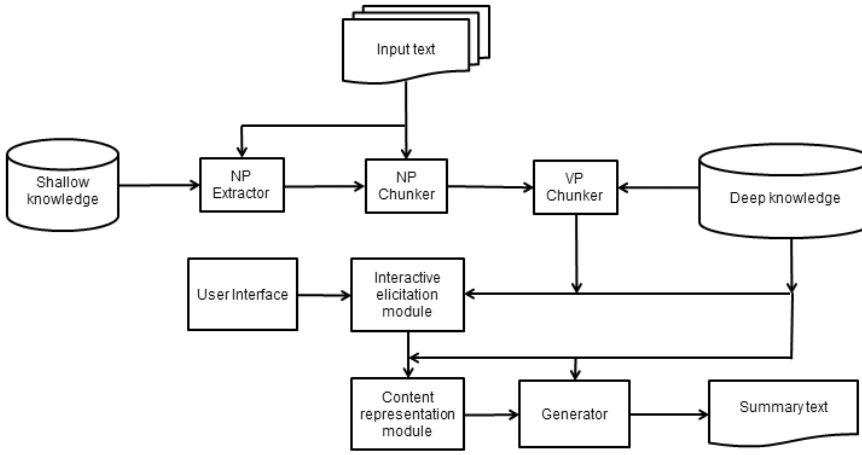


Fig. 1. An overall architecture of the summarizer

4. Knowledge

The summarizer knowledge builds on mathematical modelling corpus analysis and consists of application specific lexicons and processing rules. It is organized in blocks used by different system modules.

Block 1 (shallow knowledge) is used by the NP extractor and chunker. The NP extractor knowledge consists of a number of shallow lexicons and deletion rules; the lexicons are parts-of-speech-sorted lists of wordforms from the corpus that are forbidden in certain positions in Russian NPs; the rules draw on NP structural restrictions. The extractor was ported from English and tuned to the domain in question; see details in [19].

Block 2 (deep knowledge) includes an information-rich lexicon of predicates and rules to handle predicate knowledge in the analysis and generation modules of the summarizer. The lexicon is composed of a set of single sense entries defined as follows:

dictionary::= {entry}+

entry::= part-of-speech, typed morphological features
(number, gender, time, etc.),
domain relevant morphological forms explicitly listed in the entry
case-roles, linear patterns

The set of parameters (or fields) for predicate specification in the entries of the lexicon is strictly determined by the needs of application and corpus analysis. Explicitly listed relevant morphological forms of the predicates help their identification in the input text. Certain conventions are used in organizing knowledge in the lexicon

that facilitates processing. For example, passive and active forms of a predicate are considered to be realizations of different lexemes.

The maximum set of the predicate case-roles includes *subject, direct object, indirect object1, manner, place, means, purpose, time, source, destination, condition, agent* and *other.* Linear patterns code information about the order of realization of case-roles of certain semantic status in the syntactic structure of the summary sentence. The knowledge base is further augmented with content elicitation templates built on the predicate knowledge and so as to prevent generation of summary sentences with complex syntactic structures. The templates are empty predicate—argument frames. Each slot in the elicitation template contains a main (predicate) slot and slots corresponding to the particular predicate case-roles. The domain restrictions are reflected in the corpus-based predicate vocabulary and predicate morphological, syntactic and semantic properties (case-role sets and patterns).

5. Workflow

The workflow in the summarizer is as follows:

Input markup. The goals of this automatic analysis stage is to (i) alert the user on the paper terminology, (ii) automate user manipulations with the text chunks and (iii) link predicate lexemes with the deep knowledge in the predicate lexicon, elicitation templates and processing rules in the system knowledge base. The document first goes to the automatic extractor, which produces a list of one- to four component noun phrases. The output NP list is then supplied to a shallow analyzer component, which by matching the extracted list against the input text from left to right chunks NP terminology in the document. The remaining text is supplied to another analyzer component, which matches it against the strings in the morphological zones of the predicate lexicon. In case of a match the text string is chunked as a predicate and linked to all the information of the corresponding predicate entry. The fact that the predicate chunker does not run within chunked NPs practically lifts the ambiguity problem in predicate identification. Finally the document is turned into an interactive ("clickable") text with NPs and VPs highlighted and presented to the user.

Morphological normalization of text predicates. The goal of this stage is to create a menu of predicates in normalized forms to facilitate content selection. In our system we normalize text forms of the predicates to their finite forms while keeping the features of time, voice, aspect, gender and number of these lexemes as used in the text. For example, if the text form of a predicate is "*разделяющего* " ("*separating*") with the morphological features of *present participle, masculine, singular, genitive, active voice*, it is stored in the predicate menu in the form of "*разделяет*" with morphological features of *active finite form, present tense, masculine, singular, genitive.* This makes it easy for the user to see the lexeme as a syntactic predicate of a summary sentence. Predicate normalization rules work over the knowledge stored in the morphological zone of the lexicon.

---

[1] These are just case-role labels to show that their fillers occupy positions of subject, direct and indirect objects in the syntactic structure of a text sentence.

Elicitation of the summary content. One goal of this interactive stage is to elicit from the user the knowledge about the summary content while controlling the syntactic structure of the nascent summary. Another goal of this stage is to facilitate building internal representation of the elicited content. In fact, the user is implicitly prompted to decompose a complex input text into predicate phrases by supplying text strings into predicate templates from the system knowledge base. The elicitation procedure halts when all the content for the summary is elicited.

Internal content representation. Once the user fills the appropriate slots of a predicate template, the system automatically produces the internal representation of an elicited quantum of the summary content as follows

(2) *(**P 2 "(V_pres_3prs_sg) "является** / is"*
***(1 subj "данный алгоритм** /the given algorithm")*
***(4indir-obj "основной частью итерационного метода решения задачи сильной сходимости** /main part of the iteration method for solving the strong convergence problem")*
***(5place "в распознавании образов"** / in image recognition))]*

The output of this stage is a set of filled predicate templates.

Generation of the summary text. The content representation is passed to the text planner that treats individual predicate templates as representations of summary sentences, while case-role fillers are treated as sentence constituents. By default the planner orders a set of predicate-argument structures following the order of their creation by the user. This order can be changed interactively through the user interface. The order of case-role fillers as sentence constituents (that are treated as blocks without any analysis) is defined according to the corresponding linear pattern in the predicate entry in the lexicon. For example, the internal representation of the Russian predicate P2 *"является/is"* shown in (2) will be linearized according to the pattern (5 1 × 4) stored in the lexicon entry of this predicate:

(3) ***(**5place**: "в распознавании образов"** /in image recognition) **(**1 subj **"данный алгоритм** /the given algorithm") **x : "является́ся** / is" **(**4indir-obj **"основной частью итерационного метода решения задачи сильной сходимости** / the main part of the iteration method for solving the strong convergence problem")*

6. Implementation

The summarization methodology is implemented it into a program,—a text-to-text interactive summarizer for the domain of scientific papers on mathematical modelling in the Russian language. The programming is done in C++ for the Windows operational environment. Using common graphical user interface tools (such as dialogue boxes, menus, templates, etc.), the system guides the user through the paces of content selection and creation of a highly readable and translatable summary.

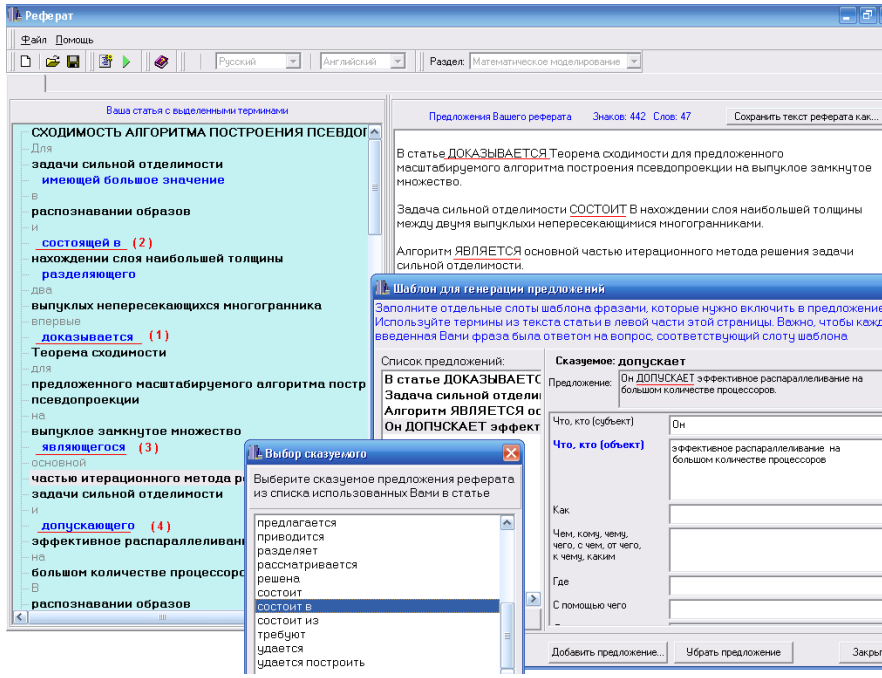


Fig. 2. A screenshot of a fragment of the summarizer interface in the process of content elicitation

The start up page of the interface displays two empty panes. On downloading a document the interface automatically displays the document text made interactive with noun and predicate phrases highlighted in different colours as shown in Fig. 2. Simultaneously the first elicitation template pops up completely or partially filled. The predicate slot is filled with the lexeme *"рассматривается/considered"*; the "subject" slot is filled with the most relevant term (the first NP from the title) and the "place" slot contains a fixed string *"в статье/in the paper"*. The user can accept, edit or delete the template or any of its fillers. Any text string can be transferred from the left pane into an elicitation template on a double click. Template slots that correspond to predicate case-roles are marked with "human" questions conveying the semantic status of the case-roles/slots (e.g.," 5 place"à"где"/"where"; " 3 manner" à"где"/"how"). The user is supposed to fill the slots with text strings that answer the questions. Immediately on the completion of filling a predicate template a summary sentence is generated on the back right pane for user control as shown in Fig. 2. The "Add a new sentence" button on the predicate template calls a predicate menu. A predicate selected in menu calls for the corresponding template to be filled. The pop-up predicate menu is provided with an empty type in area in case the user prefers to use a predicate other than one of those used in the document. A "new" typed in predicate calls a default predicate template. It is possible to call for a predicate template using a "short cut" by directly clicking on the highlighted predicate in interactive text. The Interface includes a spell checker and provides supporting information such as the number of characters

and words of already generated sentences (shown on the top of the right pane). The "Save as text" button saves a summary in a text file. The "Save project" and "Open project" selections in the main File menu allow for saving/opening summary drafts to work on them in several takes.

7. Evaluation

The specificity of the approach is reflected in the aspects that undergone evaluation. Neither content selection, nor document compression evaluations are applicable. The content as selected by the user was considered to be fully correct. The level of document compression fully depends on the lengths of text chunks that the user supplies to predicate slots. The generator is responsible for the predicate morphological forms that should agree with the fillers in the subject slots and the order of realization of case-role slots fillers that are treated as text blocs without any change.

The evaluation was therefore done on the following four levels: grammaticality, readability, translatability and user satisfaction based on the expert judgment. The body of experts included translators, professors, and students from mathematical and linguistic departments of the South Ural State University, Russia (http://www.susu.ac.ru). All experts were native Russian speakers with good proficiency in English. The evaluation tasks were divided between the participating experts. As a starting point 200 full papers were divided between the participating experts—the mathematicians who were first asked to create summaries following instructions only, then with the help of existing on-line style checkers and at last they created summaries with the tool (to guarantee correct content). The on-line Russian style checkers were almost immediately rejected by the experts as completely useless for the task. Instructions-only summaries were also rejected by mathematicians as too "linguistic" or vice-versa 'too general" to help summary composition.

Grammaticality of the tool generated summaries was judged by the students—linguists. The sentences of the summary were considered grammatical if they obeyed the rules of the Russian grammar. The experts reported ~ 91% of grammatically correct sentences. Grammar mistakes were caused by the incomplete coverage of the predicate lexicon, when due to the lack of a predicate default templates (and linear patterns) were used which caused problems in word order.

Improvement in readability of the tool-generated summaries was judged by participating mathematicians and translators by comparing them with the original (author written summitries) ones. For translators (who are not mathematicians) readability means clear and unambiguous syntax. This evaluation was qualitative and both parties reported improvement in this parameter.

To evaluate translatability translators and students-linguists were asked to translate original and tool generated summaries into English with two most popular and best developed for Russian free online MT systems,—PROMT and GOOGLE TRANSLATE. The former is a rule-based MT system, while the latter is a statistical MT system. Terminology translation was excluded from the examination, as none of these online systems were supposed to cover properly the terminology of our domain. Syntactic

mistakes that occurred due to wrong terminology translation were little in number and, therefore, neglected. First of all it was found that both RBMT (PROMT) and SMT (GOOGLE TRANSLATE) suffer, though differently, from the same linguistic phenomena in the source Russian texts. There was practically no author written summary which could provide for a correct MT. On the other hand, practically all tool generated summaries (terminology neglected) provided for correct MT.

Two groups of experts, who have worked with the system, authors (mathematicians) and translators reported high satisfaction with the summarization results and the simplicity of the knowledge elicitation procedure. It took around an hour of training for them to get acquainted with the system. They underline the usefulness of full document mark-up for the summarization procedure that made the input much more understandable. The quality of mark-up depends on the correctness of NP and VP chucking. The latter was evaluated by comparing tool-chunked NPs and VPs with the gold lists of these phrases (manually for every document). The results proved to be rather high due to the high quality performance of the NP extractor, rich morphology of the Russian language and tuning the system knowledge to the restricted domain (see Table 1).

Table 1. Chunking evaluation results

NP chunking (%)		**Predicate chunking (%)**		**Grammar (%)**
recall	precision	recall	precision	correctness
95	94	98	96	93

8. Conclusions

In this paper we described a methodology for developing a system for generating a single document text-to-text-summary that involves computer-supported human intervention at the stage of content selection and provides for high readability and translatability of the summary. The validity of the approach was proved by its implementation into a summarizer for scientific papers in the domain of mathematical modelling in the Russian language. The summarizer is fully operational.

The static knowledge sources including the shallow and deep lexicons as well as other analysis- and transfer-related knowledge blocks have been compiled based on the sublanguage analysis and provide for good coverage. The summarization algorithm is universal and robust as it excludes such statistically or NLP expensive techniques as combinatorial computations or tagging and parsing. The evaluation results presented in this paper confirm the viability of the approach.

We plan to extend this work in a number of ways. We are currently working on making our summarizer fully automatic.

Due to the universal metalanguage of knowledge representation and language-independent processing algorithms the methodology presented in this paper is highly portable and allows for extending the summarizer on other domains and languages. We intend just that. It takes tuning the shallow lexicons of the NP extractor to a new domain that within

one language can cost a couple of one man days and update a lexicon of predicates, a lot of predicate knowledge already acquired being reused. The program shell and developer tools are completely reusable. One more perspective is to apply the techniques described to multilingual applications, e.g., multilingual search or machine translation.

References

1. *Alekseev A., Loukachevitch N.* (2012), Use of Multiple Features for Extracting Topics from News Clusters. Proc SYRCODIS'2012, pp. 3–11.
2. *Barzilay R. and Lillian Lee.* (2004), Catching the drift: Probabilistic content models, with applications to generation and summarization. In DanielMarcu Susan Dumais and Salim Roukos, editors, HLT-NAACL 2004: Main Proceedings, pages 113–120, Boston.
3. *Barzilay R. and Kathleen R. McKeown.* (2005), Sentence fusion for multidocument news summarization. Computational Linguistics, 31(3): 297–328.
4. *Clarke J. and Mirella Lapata.* (2007), Modelling compression with discourse constraints. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), pages 1–11.
5. *Daumé H. III and Daniel Marcu.* (2005), Induction of word and phrase alignments for automatic document summarization. Computational Linguistics, 31(4): 505–530, December.
6. *Fathy Ibragim.* (2012), Rich semantic representation based approach for text generation. In Proc. The 8th International Conference on Informatics and Systems (INFOS). Cairo, Egypt 14–16 May.
7. *Hahn U., and Reimer U.* (2001), Knowledge-Based Text Summarization: Salience and Generation Operators for knowledge Base Abstraction. In Inderjeet Mani and Mark T. Maybury, editors, Advances in Automatic Text Summarization! Massacusets Institute of Technology, pp. 215–232.
8. *Khodra M. L., D. H. Widyantoro, E. A. Aziz, B. R. Trilaksono.* (2011), Free Model of Sentence Classifier for Automatic Extraction of Topic Sentences. Journal of ICT Research and Applications, Vol. 5C No. 1.
9. *Jinha A. E.* (2010), Article 50 Million: An Estimate of the Number of Scholarly Articles in Existence. Learned Publishing, 23 (3) (2010), pp. 258–263.
10. *Lapata M.*(2003), Probabilistic text structuring: Experiments with sentence ordering. In Proceedings of the 41st AnnualMeeting of the Association for Computational Linguistics, pages 545–552, Sapporo, Japan.
11. *Leontyeva N. N.* (2003), Semantic Dictionary for Text Understanding and Summarization // International Journal of Translation. V. 15. № 1. P. 107–114.
12. *Leuski A., Lin C-Y., Hovy E.* (2003), iNeATS: Interactive Multi-Document Summarization. In Proc. The 41st AnnualMeeting of the Association for Computational Linguistics, Sapporo, Japan.

13. *Lin, J., M. Nitin., B. Dorr.* (2010), Putting the User in the Loop: Interactive Maximal Marginal Relevance for Query-Focused Summarization. In Proceedings of the The 48th Annual Meeting of the Association for Computational Linguistics. Uppsala, Sweden, July 11–16.
14. *Lloret E. A.* (2009), Gradual Combination of Features for Building Automatic Summarisation Systems Text [Text] / E. Lloret, M. Palomar // Speech and Dialogue. — Heidelberg, — P. 16–23.
15. *Lukashevich N. V., Dobrov B. V.* (2009), Automatic annotation of a news cluster based on thematic representation [Avtomaticheskoye annotirovaniye novostnogo klastera na osnove tematicheskogo predstavleniya], Computational Linguistics and Intelligent Technologies: Proc. The International Conference Dialog'2009, [Komp'yuternaya lingvistika i intellektual'nyye tekhnologii: trudy Mezhdunarodnoy konferentsii Dialog'2009] Vol. 8 (15). c. 299–305.
16. *McKeown K. R.* (1985), Text Generation: Using Discourse Strategies and Focus Constraints to Generate Natural Language Text. Cambridge University Press.
17. *Saggion, H. A.* (2009), Classification algorithm for predicting the structure of summaries. Proc. The 2009 Workshop on Language Generation and Summarisation, ACL-IJCNLPь 2009. — Suntec, P. 31–38.
18. *Sheremetyeva S.* (2005), Embedding MT for Generation in English from a Multilingual Sheremetyeva S. 2005. Embedding MT Generation in English from a Multilingual Interface. Proceedings of the workshop on Patent Translation in conjunction with MT Summit X, Phuket, Thailand, September, 12–16.
19. *Sheremetyeva S.* (2012), Automatic Extraction of Linguistic Resources in Multiple Languages. In Proceedings of NLPCS 2012, 9th International Workshop on Natural Language Processing and Cognitive Science in conjunction with ICEIS 2012, Wroclaw, Poland.
20. *Thiago A., H. Rino and M. Nunes.* (2003), GistSumm: a summarization tool based on a new extractive method. In Proceedings of the 6th international conference on Computational processing of the Portuguese language, pages 210–218.
21. *Underwood N. L. and Jongejan B.* (2001), Translatability Checker: A Tool to Help Decide Whether to Use MT. Proceedings of MT Summit VIII, Santiago de Compostela, Spain.
22. *Wan S., Robert Dale Mark Dras, and C´ecile Paris.* (2005), Towards statistical paraphrase generation: preliminary evaluations of grammaticality. Proc. The 3rd International Workshop on Paraphrasing (IWP2005), Jeju Island, South Korea, pp. 88–95.
23. *Wan S., Robert Dale Mark Dras, and C´ecile Paris.*(2008), Seed and Grow: Augmenting Statistically Generated Summary Sentences using Schematic Word Patterns. In Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pages 543–552, Honolulu, October 2008.
24. *Wubben S., A. van den Bosch and E. Krahmer.* (2012), Sentence Simplification by Monolingual Machine Translation. In Proceedings of the 50th Annual Meeting of the Association for Computational linguistics. Jeju, Korea, July.

РУССКИЕ ЛИНГВОСПЕЦИФИЧНЫЕ ЛЕКСИЧЕСКИЕ ЕДИНИЦЫ В ПАРАЛЛЕЛЬНЫХ КОРПУСАХ: ВОЗМОЖНОСТИ ИССЛЕДОВАНИЯ И «ПОДВОДНЫЕ КАМНИ»[1]

Шмелев А. Д. (shmelev.alexei@gmail.com)

Московский педагогический государственный университет, Москва, Россия;
Институт русского языка им. В. В. Виноградова РАН, Москва, Россия

Ключевые слова: перевод, параллельный корпус, лексическая единица, семантическое различие, лингвоспецифичность, «непереводимость»

RUSSIAN LANGUAGE-SPECIFIC LEXICAL UNITS IN PARALLEL CORPORA: PROSPECTS OF INVESTIGATION AND “PITFALLS”

Shmelev A. D. (shmelev.alexei@gmail.com)

Moscow Pedagogical State University, Moscow, Russia;
Vinogradov Institute of Russian Language, Russian Academy of Sciences, Moscow, Russia

The paper deals with language-specific lexical units as they appear in parallel corpora and the degrees of linguistic specificity. It discusses new insights into the languages compared that parallel corpora can provide as well as various pitfalls on the way to an accurate account of typological and cultural differences and similarities.

In particular, it deals with Russian language-specific words, which defy translation into other languages. On the other hand, Russian language-specific words quite often appear in translations into Russian even though no exact equivalent exists in the language of the original text; special

[1] Статья написана при финансовой поддержке РФФИ (в рамках научно-исследовательского проекта РФФИ «Контрастивное корпусное исследование специфических черт семантической системы русского языка», проект № 13-06-00403 А).

attention is given to particles. The lack of such a particle where the communicative situation calls for it every so often gives the impression that we deal with a word-for-word translation of the original text containing no similar marker. In the absence of the relevant particle, the wrong implicatures may appear, or the text may cease to have coherence, or the utterance is perceived as a manifestation of arch use of language.

Key words: translation, parallel corpus, lexical unit, semantic difference, language specificity, “untranslatability”

1. Лингвоспецифичность и «переводимость»

Лингвоспецифичными называются явления, которые присутствуют не во всех языках мира; они противопоставляются явлениям, присутствующим во всех языках, или универсальным. Понятно, что установить лингвоспецифичность какого-либо языкового явления существенно проще, нежели установить универсальность: довольно привести пример хотя бы одного языка, в котором данное явление не имеет места (тогда как для доказательства универсальности, вообще говоря, следовало бы осуществить проверку по всем языкам мира).

Поскольку для установления лингвоспецифичности достаточно сопоставления с каким-то одним языком, на практике таким сопоставлением иногда и ограничиваются. В то же время сопоставление с большим числом языков дает возможность дать количественную оценку лингвоспецифичности: чем больше языков, в которых данное явление отсутствует, тем более специфично оно для языка, в котором оно имеется.

Лингвоспецифичность лексических единиц (слов в конкретном лексическом значении и фразеологизмов) заключается в специфичности их содержательной стороны (включая коннотации, фоновые компоненты значения и т. д.). Тем самым появляется почва для другой разновидности количественной оценки лингвоспецифичности: чем более своеобразна семантическая конфигурация, кроящаяся за лексической единицей, тем более лингвоспецифичной она может считаться.

Поскольку лингвоспецифичность лексической единицы определяется отсутствием у нее семантического эквивалента в языке, с которым ведется сопоставление, лингвоспецифичность естественно связывается с «непереводимостью». К сожалению, само слово «непереводимость» может ввести в заблуждение и повести к недоразумениям. Как известно, объектом практического перевода является не отдельная лексическая единица, а тот или иной текст (в котором могут содержаться такие единицы). Если иметь в виду реальную переводческую практику, любой текст оказывается «переводимым» с той степенью точности и адекватности, которая диктуется целями перевода и способностями переводчика. Поэтому точнее говорить не о «непереводимости», а об отсутствии точного словарного эквивалента в языке, который служит объектом сопоставления. Чем выше степень лингвоспецифичности лексической единицы, т. е. чем более своеобразна содержащаяся в ней семантическая

конфигурация, тем труднее найти для нее хотя бы приблизительный аналог в языке, с которым проводится сопоставление.

С этим связано то, что в реально существующих переводах для единиц с высокой степенью лингвоспецифичности часто предлагается значительное число переводных вариантов. В самом деле, если нет общепризнанного словарного аналога, выбор перевода обычно определяется специальным решением переводчика, опирающемся на его опыт, интуицию и знание обоих языков (языка оригинала и языка перевода); при этом решения обычно оказываются разными для разных контекстов или даже для одного и того же контекста у разных переводчиков. Только наивностью и непониманием сути дела можно объяснить утверждения, согласно которым то или иное слово не является лингвоспецифичным, поскольку для него в языке перевода обнаруживается множество переводных «эквивалентов»[2]; напротив того, именно обилие «эквивалентов» (а вернее — неточных аналогов) является надежным свидетельством лингвоспецифичности.

На основе сказанного возникает идея введения еще одного параметра количественной оценки лингвоспецифичности: чем более разнообразны переводы лексической единицы в реально существующих переводных текстах, тем выше ее лингвоспецифичность. Для оценки лингвоспецифичности по данному параметру можно опираться на данные какого-либо параллельного корпуса.

Однако здесь следует считаться с тем, что не всегда лингвоспецифичность ведет к разбросу переводов. Во-первых, есть заведомо лингвоспецифичные единицы, которые при переводе обычно вообще опускаются. Так, едва ли подлежит сомнению лингвоспецифичность русской частицы *же*, которую переводчики чаще всего просто опускают. В частности, это касается подавляющего большинства примеров использования этой единицы в оригинальных русских текстах, вошедших в параллельный английский подкорпус «Национального корпуса русского языка» (далее — НКРЯ)[3]. Приведем несколько характерных примеров[4]:

> — И что же, по вашему мнению, является самым важным открытием за все эти тринадцать лет? [А. Н. Стругацкий, Б. Н. Стругацкий] — “And what, in your opinion, has been the most important discovery in these thirty years?” [Antonina W. Bouis]

---

[2] Так, мне встретились подобные утверждения, отрицающие лингвоспецифичность русских слов *пошлость* и *авось*. Лингвоспецифичность прилагательного *пошлый* отвергалась на том основании, что для него «подобрать эквивалент в рамках текста переводчику несложно» (предлагались такие слова, как *kitschig*, *ordinär*, *anzüglich*, *schlüpfrig*, *vulgär*). Аргументация относительно слова *авось* была сходной: говорилось, что «гресловутый “русский авось” переводится самыми разными способами: *auf gut Gück*, *wenn was ist*, *für den Fall der Fälle*, *aufs Geratewohl*, *ins Blaue hinein*, *planlos*, *ohne Plan*, *auf Gutdünken* Идея “авося” в немецком дискурсе частотна чрезвычайно». Разбор этих и подобных утверждений содержится в моей статье [Шмелев 2014].

[3] Разумеется, следует отсеять все примеры использования этой частицы в качестве показателя тождества (т. е. в таких сочетаниях, как *тот же*, *такой же* и т. п.).

[4] Здесь и далее в целях экономии места опускаются указания на название произведения, года его создания, год создания перевода. Указываются лишь авторы и переводчики.

> Здорово, — он мне говорит. — Тебя, — говорит, — рыжий, по всему институту ищут… Тут я его так вежливенько прерываю: — Я тебе не рыжий, — говорю. — Ты мне в приятели не набивайся, шведская оглобля. — Господи, рыжий! — Говорит он в изумлении. — Да тебя же все так зовут. [А. Н. Стругацкий, Б. Н. Стругацкий] — "Fine thing," he said to me. "They're looking for you all over the institute, Red." I interrupted him right there, polite-like. "I'm not Red to you," I said. "Don't try that palsy-walsy stuff on me, you Swedish dolt." "God, Red! Everybody calls you that." [Antonina W. Bouis][5]

> Куда же мне уходить? [Н. Н. Носов)] — "Where shall I go?" [Margaret Wettlin)]

> Необходимо добавить, что на поэта иностранец с первых же слов произвел отвратительное впечатление, а Берлиозу скорее понравился… [М. А. Булгаков] — It must be added that from his first words the foreigner made a repellent impression on the poet, but Berlioz rather liked him… [Richard Pevear , Larissa Volokhonsky]

Кроме того, для многих лингвоспецифичных слов имеется принятый переводческий «эквивалент», который, не будучи семантически тождествен исходному слову, используется в подавляющем большинстве практических переводов. Скажем, выражения *утречком*, *с утра*, *под утро*, *поутру*, *под утро*[6] вполне успешно переводятся на английский язык выражением *in the morning*. Некоторая дополнительная информация, содержащаяся в данных выражениях, обычно не учитывается переводчиками (или предполагается восстанавливаемой из контекста), напр.:

> с утра еще сделав себе расписание дня [Л. Н. Толстой] — when mentally sketching out the day in the morning [Constance Garnett)]

> Анна с утра оживленно принялась за приготовление к отъезду. [Л. Н. Толстой] — Anna set eagerly to work in the morning preparing for their departure. [Constance Garnett]

> — Утречком поехали бы, разлюбезное дело, — подтверждал старик. [Л. Н. Толстой] — 'You could go on in the morning and it would be pleasanter,' said the old man, confirming what his wife had said. [Louise and Aylmer Maude]

> С утра шел тихий, без ветра, теплый дождичек…. [Л. Н. Толстой] — There was no wind; a soft warm rain had begun falling in the morning… William E. Smith]

---

[5] Обращает на себя внимание, что переводчица, по-видимому, приняла приветствие *Здоро́во* за утверждение *Здо́рово* 'очень хорошо'.

[6] Эти выражения рассматриваются в нашей статье [Зализняк, Шмелев 1997].

...он с утра поехал на Васильевский остров к Шустовой.
[Л. Н. Толстой] — ...he went in the morning to the Vasilievski Ostrov to see Shustova. [William E. Smith]

Судить начали с утра…. [А. П. Чехов] — The trial began in the morning… [Constance Garnett]

Под утро надо было идти в церковь к утрене. [А. П. Чехов] — In the morning he had to go to church to matins. [Constance Garnett]

Что за дом у нас такой! И этот с утра пьяный. [М. А. Булгаков] — 'What a house we've got… Here's this one drunk in the morning…' [Richard Pevear , Larissa Volokhonsky]

— Ну, Киса, — заметил Остап, — придется с утра сесть за работу.
[И. А. Ильф, Е. П. Петров] — "Well, Pussy," declared Ostap, "we'll have to get down to work in the morning." [John Richardson]

Тем самым невозможно прямолинейное использование измерения лингвоспецифичности, основанное на разбросе переводов. Однако остается в силе общая закономерность: разнообразие переводов лексической единицы часто свидетельствует о том, что есть основания считать ее лингвоспецифичной.

2. Лингвоспецифичные слова в параллельном корпусе: возможности и ограничения

Если мы будем ориентироваться на реально встретившиеся переводы лингвоспецифичных слов языка оригинала, мы неизбежно столкнемся еще с рядом трудностей.

Прежде всего, мы должны считаться с неизбежной на данном этапе ограниченностью объема параллельных корпусов. Утверждения, которые делаются на основании этого ограниченного объема, с большой долей вероятности опровергаются при использовании параллельного корпуса большего объема.

Так, в интересном исследовании [Добровольский 2009] отмечается, что ни в одном из контекстов с обращением *брат* или *братец*, зафиксированных в русско-английском подкорпусе НКРЯ, буквальный перевод не встретился. На этом основании высказывается предположение, что «употребление слова *brother* в качестве обращения (когда адресат не является братом говорящего или членом монашеского ордена), видимо, воспринимается как полное нарушение коммуникативной нормы». При этом отмечается, что «это предположение нуждается в дальнейшей проверке» [Добровольский 2009: 398].

Вообще говоря, и без дополнительной проверки это предположение вызывает сомнения, поскольку в английском языке существует распространенное в некоторых социальных слоях обращение *bro* (сокращенное *brother*), а также

архаизующее *brethren*; более того в оригинальных англоязычных текстах легко можно обнаружить обращение *brother* к лицам, не являющимся братом говорящего или членом монашеского ордена. Заметим, что с момента написания статьи [Добровольский 2009] объем английского подкорпуса НКРЯ существенно увеличился, так что теперь легко обнаруживаются примеры, в которых обращения *брат* или *братец* переведены как *brother*, напр.:

> Но у последнего подлюки, каков он ни есть, хоть весь извалялся он в саже и в поклонничестве, есть и у того, братцы, крупица русского чувства. [Н. В. Гоголь] — But the very meanest of these vile men, whoever he may be, given over though he be to vileness and slavishness, even he, brothers, has some grains of Russian feeling. [Isabel F. Hapgood]

> Пришел Михеев, послушал их речи и говорит: — Грех вы, братцы, великий задумали. [Л. Н. Толстой] — "Brethren," said he, "you are contemplating a grievous sin." [Louise and Aylmer Maude]

> — Здравствуйте, братцы! — сказал он. [А. П. Чехов] — "Good-day, brothers," he said. [Constance Garnett]

> Пристав поднял глаза на Осипа и спросил: — Почему же это, братец? [А. П. Чехов] — The police inspector raised his eyes to Osip and asked: "Why is this, brother"? [Constance Garnett]

> Иван Николаевич поднял свечу и вскричал: — Братья по литературе! [М. А. Булгаков] — Here Ivan Nikolaevich raised the candle and cried out: 'Brethren in literature!' [Richard Pevear, Larissa Volokhonsky]

Но дело не только и не столько в уточнении сделанного предположения. Самое беглое рассмотрение переводов русских лингвоспецифичных обращений *брат*, *братец* и *браток* на английский язык даже на том ограниченном материале, который нам предоставляет НКРЯ, позволяет видеть, сколь велика роль индивидуальных предпочтений переводчика. Некоторые переводчики предпочитают переводить обращение *братец* как *brother*, некоторые — последовательно переводят его как *my friend*, и эти предпочтения, по-видимому, оказываются более существенными, чем любые другие факторы, которые могут повлиять на выбор переводного эквивалента. При этом основанием для переводческого решения «могут оказаться поверхностное понимание, влияние переводных словарей или даже желание придать тексту перевода некоторый налет „иностранности", чтобы текст воспринимался именно как перевод» [Михайлов 2005: 381]. Поэтому, когда русский язык выступает в качестве языка оригинала, данные параллельного корпуса дают нам заведомо недостаточную информацию о семантических особенностях русских лингвоспецифичных лексических единиц.

С другой стороны, параллельный корпус может дать весьма ценные данные при рассмотрении лингвоспецифичных слов в языке перевода. Применительно

к русским лингвоспецифичным словам это означает, что для проникновения в семантические секреты русских лингвоспецифичных слов более показательным может оказаться анализ переводов не с русского языка, а на русский язык. В самом деле, если лингвоспецифичное слово появляется в переводе, закономерным оказывается вопрос, какие особенности оригинала побуждают переводчика употребить это слово.

В статье [Levontina, Shmelev 2005] мы иллюстрировали похожее соображение на материале русской частицы *еще* в некотором особом типе употребления. Мы отметили, что фраза Терминатора *I'll be back* естественно может быть переведена на русский язык как *Я еще вернусь* (а не просто *Я вернусь*) и попытались описать условия появления частицы *еще* в контекстах такого рода.

Можно привести похожий пример. В книге Лео Ростена *The Joys of Yiddish* приводится следующая история:

> Mr. Sokoloff has had dinner for twenty years in the same restaurant on the Second Avenue. This evening, as always, he orders bouillon. The waiter brings it, and wants to go back, but Mr. Sokoloff addresses him: "Waiter!" — "Yes, please?" — "Be so kind to taste this soup." — "But Mr. Sokoloff, you have come here for twenty years and you have never complained." — "Please", repeats Mr. Sokoloff obstinately, "taste this soup." — "But what is the matter, Mr. Sokoloff?" — "Please taste it." — "All right", the waiter says. "But… a moment. Where is the spoon?" — *"Aha!",* says Mr. Sokoloff.

Почти все, кто прочел эту историю и хочет рассказать ее на русском языке, передают последние фразы официанта следующим образом: *Но… минуточку. Где же ложка?* Частица *же* почти неизбежно появляется в переводе, и возникает вопрос, какие особенности оригинала заставляют ее использовать. Ответ на этот вопрос позволяет приблизиться к пониманию условий употребления частицы *же* в русском языке.

Анализ появления лингвоспецифичных слов в русском переводе может использоваться как средство проверки их семантического анализа. Так для частиц *разве* и *неужели* в работах [Булыгина, Шмелев 1982; 1987] было дано следующее описание:

- *Разве* ***p****?*
  - раньше я думал, что не ***p***
  - теперь я вижу или слышу нечто такое, что не может быть правдой если не ***p***
  - я хочу, чтобы ты сказал мне, ***p*** или не ***p***
- *Неужели* ***p****?*
  - я думал, что ***p*** невозможно
  - некоторые вещи заставляют меня думать, что, может быть, ***p***
  - я говорю тебе, что мне трудно поверить, что ***p***
  - я хочу, чтобы ты сказал мне, ***p*** или не ***p***

Материал английского подкорпуса НКРЯ подтверждает это описание:

> The agent pulled the car to a stop and pointed between two fountains to a large door in the side of the pyramid. “There is the entrance. Good luck, monsieur.” “You’re not coming?” “My orders are to leave you here. I have other business to attend to.” [Dan Brown] — Агент остановил машину и указал на большую дверь в пирамиде между двух фонтанов. — Вход там. Желаю удачи, месье. — А вы разве не со мной? — Согласно приказу я должен оставить вас здесь. У меня есть другие дела. [Н. Рейн]

Мы имеем:

- ***p*** = ‘Агент не пойдет с Лэнгдоном’
- Лэнгдон думал, что агент пойдет с ним (т. е. что не ***p***)
- The agent said, “Good luck”, which implied that he was leaving (that is, not coming with Langdon)

В НКРЯ обнаруживается множество других примеров из НКРЯ, в которых в русском переводе употреблена частица *разве*, и все они укладываются в предложенное описание.

Частица *неужели* указывает на малую априорную вероятность *p* (с точки зрения говорящего), и примеры из НКРЯ подтверждают это (в них *неужели* в переводе используется в соответствии с выражениями *I can’t believe*, *Do you honestly think*, вопросами типа *Can it be…?* и т. п.)

Употребление частиц *разве* и *неужели* с отрицанием в целом в целом также вкладывается в предложенное описание. Возможно следующая детализация. Вопросы *Разве нет? Разве не так?* представляют собою не реакцию на высказывание собеседника, а своего рода tag questions. Их смысл — уточнить у собеседника, согласен ли он со сказанным. Вопрос *Неужели нет?* представляет собою типичный вопрос-реакцию. Его смысл — собеседник спросил о чем-то, что не вызывает у говорящего сомнения, и говорящий говорит: «Конечно! Как можно в этом сомневаться?»

Вопросы *Разве нет? Разве не так?* в переводах подтверждают предложенное описание: они служат переводами разнообразных tag questions. Вопрос *Неужели нет?* в переводах, представленных в НКРЯ, не встречается.

Русские лингвоспецифичные выражения *плюнуть/плевать* и *махнуть рукой* указывают на то, что безразличие субъекта. В соответствии с этим оборот *(мне) плевать* регулярно появляется в переводах в соответствии с английским *I don’t care*, хотя о плевках в оригинале ничего не говорится, напр.:

> I don’t care what you say about me, just spell my name right! [Dan Brown] — Мне плевать, что вы обо мне говорите, но только произносите мое имя без ошибок! [Г. Косов]

В подавляющем большинстве примеров использования в переводе выражений *плюнуть/плевать* и *махнуть рукой* для указания на безразличие в оригинале ничего не говорилось ни о плевках и о жесте ‘махнуть рукой’:

Rémy didn't give a damn about the Grail, except that the Teacher refused to pay him until it was found. [Dan Brown] — Лично он, Реми, плевать хотел на этот Грааль, но Учитель сказал, что не заплатит ничего до тех пор, пока он не будет найден. [Н. Рейн]

She doesn't give a shit what you're doing. [Lauren Weisberger] — Она плевать на тебя хотела. [М. Маяков, Т. Шабаева]

What would it say about God if God had done nothing? That the Almighty did not care? [Dan Brown. Angels and Demons (2000)]

I didn't give a damn how I looked. [J. D. Salinger] — Плевать мне было, какой у меня вид. [Р. Райт-Ковалёва]

…she wouldn't worry about words from High Up. [J. R. R. Tolkien] — Шелоб плюет на высочайшие приказы. [М. Каменкович, В. Каррик]

…regardless of the consequences for them; they had given up and gone home. [Stephen King] — …не думая о последствиях: плюнули на все и вернулись домой. [С. Мануков]

Supposing I got like the others-not caring. [William Golding] — Ну вот возьму я и на все плюну. [Е. Суриц]

Лишь в исключительных случаях (в НКРЯ был обнаружен один пример) выражение *махнуть рукой* соответствует аналогичному жесту безразличия, на который указывал оригинальный текст:

He waved a hand like it didn't matter to him, like nothing mattered. [Michael Connelly] — Делакруа махнул рукой так, словно для него это не имело значения, словно ничто не имело значения. [Д. Вознякевич]

Мы видим, что лингвоспецифичные слова регулярно появляются в тексте перевода. Именно для таких случаев корпусная методика оказывается максимально эффективной. Появление таких слов, по-видимому, является неосознанным решением переводчика как носителя языка.

Напротив того, перевод лингвоспецифичных слов представляет собою задачу, которая, как правило, бывает отрефлектирована переводчиком. В этом отношении ценным было бы создание корпуса особого типа, в основу которого был бы положен один и тот же текст, содержащий лингвоспецифичные слова и переведенный на иностранный язык множеством разных переводчиков и снабженный их комментариями для всех трудных случаев. В этом случае мы извлекали бы из корпуса данные не столько о спонтанной языковой деятельности носителей языка, сколько об их метаязыковой рефлексии, что для многих задач не менее ценно. Однако параллельный корпус такого рода, насколько я знаю, еще не создан.

Литература

1. *Булыгина Т. В., Шмелев А. Д.* Диалогические функции некоторых типов вопросительных предложений // Известия АН СССР. Серия литературы и языка. 1982. Т. 41, № 4. С. 314–326.
2. *Булыгина Т. В., Шмелев А. Д.* О семантике частиц *разве* и *неужели* // НТИ, 1987, № 10. С. 21–25.
3. *Добровольский Д. О.* Корпус параллельных текстов в исследовании культурно-специфичной лексики // Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы. СПб.: Нестор-История, 2009. С. 383–401
4. *Зализняк Анна А., Шмелев А. Д.* Время суток и виды деятельности // Логический анализ языка. Язык и время. М., 1997. С. 229–240.
5. *Михайлов М. Н.* Частица и целое: к вопросу о поиске соответствий служебных слов в параллельном корпусе переводных текстов // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог'2005» (Звенигород, 1–6 июня, 2005 г.). М., 2005. С. 377–381.
6. *Шмелев А. Д.* (2014) Язык и культура: есть ли точки соприкосновения? // Труды Института русского языка им. В. В. Виноградова. Вып.1. Стр. 36–116.
7. *Levontina I., Shmelev A.* The particles one cannot do without // East—West Encounter: Second International Conference on Meaning ↔ Text Theory. Moscow, 2005. P. 258–267.

References

1. *Bulygina T. V., Shmelev A. D.* (1982) Dialogical functions of certain types of interrogative sentences [Dialogicheskie funktsii nekotorykh tipov voprositel'nykh predlozheniy], Proceedings of the Academy of Sciences of the USSR, Series of Language and Literature [Izvestiya AN SSSR. Seriya literatury i yazyka], Vol. 41, 4, pp. 314–326.
2. *Bulygina T. V., Shmelev A. D.* (1987) On the semantics of the particles *razve* and *neuzheli* [O semantike chastits *razve* i *neuzheli*], Science and Technical Information [Nauchno-tekhnicheskaya informatsiya], 10, pp. 21–25.
3. *Dobrovol'skiy D. O.* (2009) Parallel corpus in the investigation of culture specific lexicon [Korpus parallel'nykh tekstov v issledovanii kul'turno-spetsifichnoy leksiki], Russian National Corpus: 2006–2008. New Results and Perspectives [Natsional'nyy korpus russkogo yazyka: 2006–2008. Novye rezul'taty i perspektivy], Nestor-Istoriya, S.-Petgersburg, pp. 383–401
4. *Levontina I., Shmelev A.* The particles one cannot do without // East—West Encounter: Second International Conference on Meaning ↔ Text Theory. Moscow, 2005. P. 258–267.
5. *Mikhaylov M. N.* (2005) The particle in the text: is it possible to check correspondences of functional words in parallel corpora [Chastitsa i tseloe: k voprosu o poiske sootvetstviy sluzhebnykh slov v parallel'nom korpuse perevodnykh

tekstov], Computational Linguistics and Intellectual Technologies. Proceedings of the International Conference “Dialog 2005” [Komp’yuternaya lingvistika i intellektual’nye tekhnologii. Trudy mezhdunarodnoy konferentsii “Dialog’2005”] Moscow, pp. 377–381.

6. *Shmelev A. D.* (2014) Language and culture: do they have points of interaction? [Yazyk i kul’tura: est’ li tochki soprikosnoveniya?], Proceedings of the V. V. Vinogradov Institute of Russian Language [Trudy Instituta russkogo yazyka im. V. V. Vinogradova], 1, pp. 36–116.
7. *Zaliznyak Anna A., Shmelev A. D.* (1997) Time and Human Activities [Vremya sutok i vidy deyatel’nosti], Logical Analysis of Language. Language and Time [Logicheskiy analiz yazyka. Yazyk i vremya], Moscow, pp. 229–240.

ГЕНЕРАЦИЯ ЕСТЕСТВЕННОГО ЯЗЫКА, ПАРАФРАЗ И АВТОМАТИЧЕСКОЕ ОБОБЩЕНИЕ ОТЗЫВОВ ПОЛЬЗОВАТЕЛЕЙ С ПОМОЩЬЮ РЕКУРРЕНТНЫХ НЕЙРОННЫХ СЕТЕЙ

Тарасов Д. С. (dtarasov3@gmail.com)

Интернет-портал reviewdot.ru, Россия, Казань

Ключевые слова: генерация естественного языка, генерация парафраз, автоматическое обобщение отзывов, рекуррентные нейронные сети

NATURAL LANGUAGE GENERATION, PARAPHRASING AND SUMMARIZATION OF USER REVIEWS WITH RECURRENT NEURAL NETWORKS

Tarasov D. S. (dtarasov3@gmail.com)

ReviewDot Research, Kazan, Russia

Multi-Document summarization and sentence generation are important challenges in natural language processing. This paper presents recurrent neural network (RNN) architecture capable of producing abstractive document summaries, as well as generating novel paraphrases of input sentences in the same language. We demonstrate practical application of our system on the task of multiple consumer reviews summarization.

Keywords: natural language generation, paraphrase generation, automatic summarization of user reviews, recurrent neural networks

1. Introduction

The main role of automatic document summarization is to help readers to understand most important points of long documents without much effort. One particular area of document summarization that attracted a lot of research attention is automatic

summarization of consumer reviews, also called opinion summarization. It is traditionally based on feature selection, feature rating and identifying important sentences, leading to so called extractive summaries (summaries that consists of original sentences extracted from user reviews) [Mei et al, 2007; Liu J. et al, 2012, Liu C. et al, 2012, Raut and Londhe, 2014]. Another kind of summaries is abstractive summaries (texts that summarize essential facts mentioned in reviews without using original sentences). Such texts tend to have better coverage for a particular level of conciseness, and to be less redundant and more coherent [Carenini et al, 2006]. They also can be constructed to target particular goals, such as summarization, comparison or recommendation.

Abstractive summarizers rely on natural language generation systems, that are currently designed using a lot of expert linguistic knowledge, heuristics and complex pipelines (that typically include text planner, sentence planner and surface realizer) [Fabbrizio et al, 2014]. Therefore adapting such systems to new languages and domains can be difficult. Up until now, only a few works considered machine learning based (trainable) language generation systems, and their success was limited [Ratnaparkhi, 2000;Hammervold, 2000]. However, recent research on neural networks demonstrated their capabilities to generate novel descriptions of pictures using purely machine-learning methods [Mao et al, 2014].

In this work we explore application of similar methodology to the domain of consumer reviews. We describe and evaluate recurrent neural network (RNN) model capable of generating novel sentences and document summaries.

To achieve this, we train recurrent neural network language model on a large number of sentences describing positive and negative aspects of various consumer products. In our setup, RNN task is to predict next word given current word and additional sentence-level semantic information that include sentence polarity, sentence length, product category and bag of aspects vector. In the test phase we give RNN sentence-level features vector and generate corresponding sentence.

We demonstrate that such relatively simple model can generate novel paraphrases that capture original meaning and show that this ability can be used to “compress” multiple important points about the product in one statement, thus producing concise multi-document summary. To do this, we first compute semantic vectors for all sentences in all available user reviews of a given product, combine them into two semantic vectors—positive (containing bag of positive aspects) and negative (containing bag of negative aspects). We then feed these vectors to language-generating RNN, obtaining sentences that sum up negative and positive product sides.

2. Related work

Convolutional neural networks were used for generation of extractive summaries of movie reviews [Denil et al, 2014]. In [Iyyer, 2014] paraphrase generation using tree-based autoencoders was demonstrated, however, no evolution of paraphrase quality was presented aside from few paraphrase examples. The approach of [Iyyer, 2014] also relies on dependency parse trees. Our method in contrast, does not use sentences parsers. It can be viewed as similar to encoder-decoder machine

translation models [Cho et al, 2014], while our RNN architecture is different and inspired by method of [Mao et al, 2014] where RNN was used to generate descriptions of pictures. We are not aware of any prior application of such models to abstractive text summarization or paraphrase generation.

3. Methods and algorithms

3.1. Datasets

We use database of 820,000 consumer reviews in Russian language from reviewdot.ru that was obtained by automatic crawling of more than 200 different web-resources. From that database we selected 120,000 reviews in 15 different product categories that had three sections (positive points, negative points and comments). These three sections are commonly used in Russian consumer reviews websites and reviewdot.ru crawler automatically detects them using heuristics-based algorithm. We then exclude sentences with unknown polarity and those with length more than 25 words, resulting in 56,000 training sentences. All sentences were padded with <START> and <END> special symbols.

3.2. Summarization Recurrent neural network model

The structure of our summarization recurrent neural network (s-RNN) is shown in Figure 1. The s-RNN model is deeper than the simple RNN model and similar to multimodal RNN introduced in [Mao et al, 2014]. It has five layers in each time frame: the input word layer, one projection layer, the recurrent layer, the summarization layer, and the softmax layer.

Projection layer implements table-lookup operation, converting word to real-valued embedding vector. Embedding vectors are obtained by training recurrent neural network language model [Mikolov et al, 2010] on 30M words dataset of consumer reviews.

Recurrent layer implements standard Elman-type [Elman, 1990] recurrent function:

$$h(t) = f\left(Wx(t) + Vh(t-1) + b\right)$$

Here f is a nonlinear function, (in our case hyperbolic tangent function), W and V are weight matrices between the projection and recurrent layer, and between the hidden units. U is the output weight matrix, b is bias vector connected to hidden and output units.

After the recurrent layer, we set up a summarization layer that connects the language model part and sentence-level semantics in s-RNN model. The language model part includes the projection layer and the recurrent layer. The sentence-level semantics contains the sentence features vector. We use sentence polarity, product category, bag-of-aspect-terms vector and sentence length as sentence-level features. While it is possible to incorporate more complex features, including these learned

by unsupervised neural network models, for this proof-of-principle experiment we avoid these additional complexities.

The softmax layer on top of the network generates the probability distribution of the next word.

Our s-RNN model was trained using backpropagation through time (BPTT) [Werbos, 1990] method with mini-batch gradient descent using one sentence per mini-batch as described in [Mesnil et al, 2013].

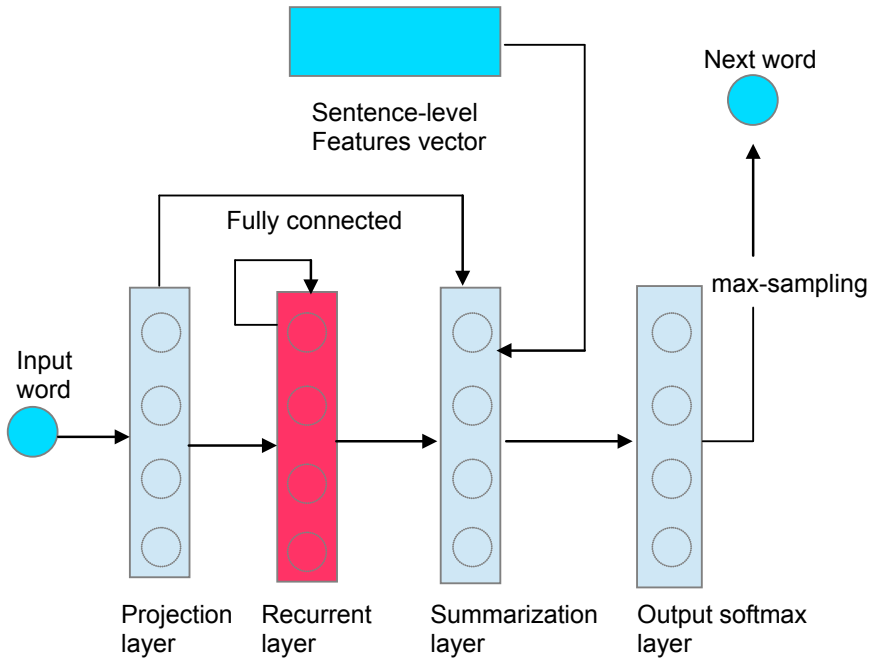


Figure 1. Architecture of summarization recurrent neural network

4. Results and discussion

4.1. Paraphrasing

To produce paraphrases, we give network sentence-level features vector of original sentence and then generate new sentence word-by-word, beginning from "<START>" symbol and stopping after network generates "<END>" symbol. Sentences quality was assessed by two human judges, who were asked to label sentences as "grammatically correct/not correct" and "conveying original meaning, not conveying original meaning". Results are summarized in table 1

Table 1. Human judgment of generated paraphrases (total of 75 sentences)

Human judgment	**Percentage of sentences (average value from results of two human judges)**
Grammatically correct and conveying original meaning	65%
Conveying original meaning but not necessary correct	78%
Correct, but not conveying original meaning	18%

Table 2. Example of original and paraphrased text

Original	**Paraphrased by neural network (two samples)**
Качественный и красивый корпус. (good and nice looking case)*	Красивый, качественный корпус, живучий. (Nice looking, good case, reliable) металический корпус довольно красиво выглядит (Metallic case looks reasonable good)
Большие кнопки с цифрами,хорошее качество изготовления,приятный корпус, лёгкий, громкий сигнал вызова (Big buttons with numbers, good production quality, nice case, lightweight, loud signal)	Вполне нравится пластик, цифры кнопок / качество изготовления металлический корпус. (Good quality of plastics, number of buttons / quality of production, metallic case) но кнопки , корпус , хорошее качество фотографий (But buttons, case, good quality of photo)
Довольно сносное качество исполнения, легкий и вполне удобный Quality is reasonable good, lightweight and convenient enough.	Маленький, легкий, удобный, хорошее качество звучания (Small, lightweight, convenient, good quality of sound) легкий , удобный , хорошее качество фотографий (lightweight, convenient, good quality of photo)
Очень простой, понятный и удобный в использовании. (Very simple, understandable and convenient in use)	Простой в использовании, удобный, недорогой. (Simple in use, convenient, not high-priced) мощный, простой, понятный аппарат удобный (powerful, simple, understandable device is convenient)

* English translations are human made, with an effort to preserve important sentence features.

As shown in table 2, most common mistakes are omissions of some original points and additions of new information that was not present in original sentence.

4.2. Language generation

Our design allows certain degree of control over the meaning of generated sentences. By choosing sentence-level features vector we can instruct the network, for example, to "say something good about screen and sound quality in about ten words".

We found, that better sentences are produced when number of words is set to roughly triple of the number of aspect terms. With smaller sentences, RNN just lists all aspects, and with larger values it tend to produce long phrases without well-defined meaning ("bright display from outside") and undesired additions such as "smart helps" (Table 3).

Table 3. Examples of sentences produced by s-RNN (polarity set to "positive" and aspects set to "battery, screen, convenience")

Desired sentence length	output
3	батарея, экран, удобный (battery, screen, convenient)
5	аккумулятор, размер дисплея солидный, эргономика (accumulator, impressive display size, ergonomics)
10	быстрый аккумулятор, яркий внешне дисплей, удобный функционал, умный помогает. (fast accumulator, bright display from outside, convenient functions, smart helps)

4.3. Summarization of multiple user reviews

Language-generating capacity of our RNN can be used for producing abstractive summaries of multiple user reviews. To achieve that we generate synthetic sentence-level feature vectors by running aspect-based sentiment analysis over all sentences of reviews subjected to summarization, using extracted aspect terms and polarities to generate feature vectors.

The major obstacle here is that our feature vectors capture only coarse-grained information (i.e. they can tell that display is good, but information why it is good is lost). Thus direct application of s-RNN usually leads to production of rather generic or plainly incorrect summaries.

To circumvent this problem, we use additional dynamic training step that consists of running one iteration of gradient descent over all sentences with aspect terms. We found that this method considerably improves quality of summaries, and allows incorporating fine-grained device-specific information.

Quality of review summaries were evaluated by two human judges who were given original reviews and asked to rate summary quality as good, acceptable or unacceptable. Table 4 presents averaged results.

Overall we found, that our method often produces summaries of reasonable quality, while still making a number of mistakes. Most commonly observed problem is inclusion

of seemingly irrelevant statements, such as "lot of different days". Also, we observed significant number of ungrammatical sentences, that can be result of relatively small training sample size, failure of RNN to capture long-term grammatical dependencies, and/or grammatical errors in the training samples (since user reviews typically contain certain number of ungrammatical phrases). The extent to which these factors contribute to generation of grammar errors is presently unknown and needs further investigation.

Still, we find it impressive that such relatively simple method can be used to solve multi-document summarization task—a problem that is generally considered difficult in natural language processing. Future work should include evaluation of proposed methods on different datasets and also investigation of possible use of trainable sentence-level feature vectors instead of pre-defined ones.

Table 4. Human evaluation of review summaries (100 summaries total).

Quality rating	Percentage of review summaries
Good	35%
Acceptable	44%
Unacceptable	21%

Table 5. Examples of generated summaries for two different mobile phones

Positives	Negatives
Качество звука, удобный интерфейс, очень долго держит заряд. Отзывчивый экран, громкий звонок, крупный шрифт, рабочий день. Приятно лежит в руках, 2 сим—карты выручают. Качество сборки, батарея, удобное меню, устойчив к воздействию воды. Явно лидируют, сочный дисплей, качество связи, плеер, фонарь. Хорошая фотокамера, динамик (Quality of sound, convenient user interface, very long battery life. Responsive screen, loud calling signal, large font, working day. Lies in hands nicely, 2 sim cards help. Quality of production, convenient menu, waterproof. Obviously leading, nice display, player, bright light. Good photo-camera, speaker).	Не обнаружено (not found)
Аккумулятор, скорость красивая. Дизайн, звук, функционал, масса разных дней хватает. Красив, несколько назад, процессор отзывчивый сенсор. Красивый экран, цветопередача. Дизайн, батарея, не тормозят, практичный. (Accumulator, speed is beautiful. Design, sound, functions, lot of different days. Beatiful, few days ago, processor, responsive sensor. Nice screen, color reproduction. Design and battery is not slow, practical).	Скользкий панель громкости тиховат. Стирается, заметно ос виснет, появляется белый экран. (Slippery panel of volume is too quiet. Noticable shabby, OS hangs and white screen appears)

Acknowledgements

Author thanks anonymous reviewers for helpful comments on earlier drafts of the manuscript.

References

1. *Carenini, G., Ng, R. T., & Pauls, A.* (2006, April). Multi-Document Summarization of Evaluative Text. In EACL.
2. *Cho, K., van Merriënboer, B., Bahdanau, D., & Bengio, Y.* (2014). On the properties of neural machine translation: Encoder-decoder approaches. arXiv preprint arXiv:1409.1259.
3. *Denil, M., Demiraj, A., Kalchbrenner, N., Blunsom, P., & de Freitas, N.* (2014). Modelling, Visualising and Summarising Documents with a Single Convolutional Neural Network. arXiv preprint arXiv:1406.3830.
4. *Elman J.* (1990). Finding structure in time. Cognitive science, 14(2):179–211.
5. *Fabbrizio D., Stent A. J., & Gaizauskas, R.* (2014). A Hybrid Approach to Multi-document Summarization of Opinions in Reviews. INLG, 54.
6. *Hammervold K.* (2000, June). Sentence generation and neural networks. In Proceedings of the first international conference on Natural language generation-Volume 14 (pp. 239–246). Association for Computational Linguistics.
7. *Iyyer M., Boyd-Graber J., Daumé H.* (2014). Generating Sentences from Semantic Vector Space Representations. NIPS Workshop on Learning Semantics.
8. *Liu C., Hsaio W.-H., Lee C.-H., Lu G., Jou E.* (2012). Movie Rating and Review Summarization in Mobile Environment. IEEE Transactions on Systems, Man, and Cybernetics-Part C: Applications and Reviews, Vol. 42, No. 3, May, pp. 397–406.
9. *Liu J., Seneff S., Zue V.* (2012). Harvesting and Summarizing User-Generated Content for Advanced Speech-Based HCI. IEEE Journal of Selected Topics in Signal Processing, Vol. 6, No. 8, pp.982–992
10. *Mao, J., Xu, W., Yang, Y., Wang, J., & Yuille, A. L.* (2014). Explain images with multimodal recurrent neural networks. arXiv preprint arXiv:1410.1090.
11. *Mei Q., Ling X., Wondra M., Su H., ZHAI C.* (2007). Topic sentiment mixture: modeling facets and opinions in weblogs. In WWW '07: Proceedings of the 16th international conference on World Wide Web. ACM, New York, NY, USA, 171–180.
12. *Mesnil, G., He, X., Deng, L. & Bengio, Y.* (2013). Investigation of recurrent neural network architectures and learning methods for spoken language understanding. In INTERSPEECH pp. 3771–3775 : ISCA.
13. *Mikolov, T., Karafiát, M., Burget, L., Cernocký, J., & Khudanpur, S.* (2010, January). Recurrent neural network based language model. In INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26–30, 2010 (pp. 1045–1048).
14. *Ratnaparkhi A.* (2000, April). Trainable methods for surface natural language generation. In Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference (pp. 194–201). Association for Computational Linguistics.
15. *Raut B., Londhe D. Survey on Opinion Mining and Summarization of User Reviews on Web* (2014). International Journal of Computer Science and Information Technologies, Vol. 5 (2), 1026–1030
16. *Werbos, P. J.* (1990). Backpropagation through time: what it does and how to do it. Proceedings of the IEEE, 78(10), pp. 1550–1560

ВЫБОР СОЮЗА *И* VS. *НО/А* ПРИ СОЧИНЕНИИ ДВУХ ПРЕДЛОЖЕНИЙ (УТОЧНЕНИЕ ПОНЯТИЙ «ОЖИДАНИЕ» И «НОРМА»)

Урысон Е. В. (uryson@gmail.com)

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

Ключевые слова: сложносочиненное предложение, соединительный союз, противительный союз, норма, ожидание, организация текста, стратегия говорящего

CONJUNCTIONS *I* 'AND' VS. *NO/A* 'BUT' BETWEEN TWO COORDINATE CLAUSES (REFINEMENT OF THE TERMS "EXPECTATION" AND "NORM")

Uryson E. V. (uryson@gmail.com)

Russian Language Institute, Russain Academy of Sciences, Moscow, Russia

The paper deals with the Russian coordinating conjunctions *i* 'and' vs. *no/a* 'but' in a compound sentence. It is common knowledge that in a sentence like "*Q, i P*" the conjunction *i* 'and' marks correspondence to a certain "norm" while in a sentence like "*Q no/a P*" an adversative conjunction 'but' marks discrepancy between "norm" and the expressed state of affairs. The problem is that in some cases both *i* 'and' and *no/a* 'but' can be used. Another problem is that in some cases the usage of these conjunctions hardly can be interpreted in the terms of "norm". I demonstrate that relevant facts can be adequately described if the basic concept of semantic interpretation is "expectation", but not "norm". Expectation can be induced by (a) common knowledge of laws of nature; (b) common notions of human life, social relations, etc.; (c) text grammar. The conjunction *i* 'and' marks correspondence to the expectation and the conjunctions *no/a* 'but' mark the cancelled expectation in all cases. But the cancelled expectation is obligatorily marked in the case (a), but not (b). As for case (c), text grammar induces merely two expectations: (c1) P has the same general "microtopic" as Q; (c2) P has the same object in the focus as Q. The propositions P and Q can have the same general "microtopic", but different objects in the focus. It is Speaker who chooses strategy for marking this or that case.

Key words: compound sentence, Russian coordinating conjunctions, norm, expectation, text grammar, Speaker's strategy

1. Объект исследования и постановка задачи

Объект работы — сложносочиненное предложение, состоящее из двух предложений, соединенных союзом *и*, *но* или *а*. Таким образом, не рассматривается случай, когда союз *и* соединяет однородные члены предложения, т. е. оформляет перечисление; ср. *Маша, Миша и Наташа поехали в Грецию*.

В традиционной русистике все три союза считаются многозначными, при том что каждый из них обладает достаточно четким семантическим «ядром». Такой подход удобен для лингвистического «портретирования» каждого союза, и мы пользуемся им в работах (Урысон 2000; 2002; 2004; 2006; 2011а). Оказалось, однако, что выбор союза в сложносочиненном предложении обусловлен, прежде всего, ядерной частью его значения (которая, разумеется, должна быть семантически согласована с контекстом), а также некоторыми требованиями к организации связного текста. Поэтому для целей данной работы оказалось возможным не делить союзы на значения.

Сложносочиненное предложение может быть частью большего текста (например, повествования или диалога) или, напротив, отдельным сообщением. Если не оговорено обратное, то предложение рассматривается как отдельное сообщение.

Выбор соединительного союза (*и)* или противительного (*но, а*) при сочинении предложений *Q* и *P* обусловлен как минимум двумя факторами: (а) семантикой самого союза (Левин 1970; Санников 1989; Lakof R. 1971; Carlson 1985; Blakemore 1987; Kitis 2000; Урысон 2011а) и (б) стратегией построения высказывания. Начнем с первого фактора.

Центральным компонентом семантики союза считается ‘соответствие/несоответствие норме’, причем норма обычно понимается как нормальная картина мира, а несоответствие норме — как «отклонение от нормальной картины мира» (Вольф 1986: 186). Если сочетание ситуаций ‘Q’ и ‘P’ соответствует норме, то это маркируется соединительным союзом (Левин 1970; Санников 1989; Урысон 2000; 2004; 2011а). Ср.

(1) *Стало тепло* (Q)*, и я снял шубу* (P).

Несоответствие сочетания ‘Q’ и ‘P’ некоторой норме оформляется противительным союзом (Левин 1970; Санников 1989; Carlson 1985; Blakemore 1987; Kitis 2000; Урысон 2004; 2006; 2011а). Ср.

(2) *Стало тепло* (Q)*, но/а я не снимал шубу* (P).

Очевидно, что представление о конкретной норме (например, о том, что в теплую погоду люди носят более легкую одежду, чем в холодную) является частью наших знаний о действительности. Таким образом, употребление выбранных союзов опирается как на знание языка (в данном случае семантики союза), так и на представления о мире.

Знания о мире не являются компонентом семантической системы языка, однако в языке существуют специальные единицы, указывающие на соответствие чего-либо норме или, наоборот, на отклонение от нормы. К числу таких

единиц относятся и сочинительные союзы *и, но/а*. Понятие нормы для описания семантики введено, по-видимому, Э. Сепиром (1944), а в московской семантической школе специально обосновано Ю. Д. Апресяном (1974).

С понятием нормы тесно связано понятие ожидания, тоже широко используемое в лингвистике. В частности, центральным компонентом семантики союзов *и*, *но/а* можно считать не 'норму', а 'ожидание', точнее: 'соответствие/несоответствие ожиданию' (см. описание союзов *если* и *хотя* в книге Падучева 2004). Заметим, что ожидание индуцируется знанием нормы, поэтому во многих случаях 'соответствие/несоответствие ожиданию' равнозначно 'соответствию/несоответствию норме'. Однако ожидание может быть индуцировано и другими факторами. Возможная природа ожидания и соотношение понятий 'ожидание' и 'норма' обсуждается ниже.

Итак, мы рассматриваем сложносочиненное предложение вида *Q, и/но <а> P.* Оно описывает некоторое положение дел, т. е. совокупность ситуаций: 'Q' и 'P'. Каждая ситуация описывается своим предложением (*Q* или *P*). Что касается сочинительного союза, то он выполняет метатекстовую функцию в смысле А. Вежбицкой (1972), т. е. помогает адресату воспринять смысл всего высказывания. Рассмотрим это подробнее.

2. Ожидание и знание законов мира: два типа нормы

Первая пропозиция Q активизирует у адресата некоторый фрагмент знания. Просодически данная пропозиция оформлена как незавершенное высказывание. Благодаря этому адресат ожидает продолжения, но не любого, а такого, которое по смыслу как-то связано с пропозицией Q. (Ср. одну из первых работ на эту тему (ван Дейк 1978); наше дальнейшее рассуждение отличается от рассуждения в этой работе.) Рассмотрим некоторые возможности.

2.1. Ожидание и объективная норма

Пропозиция Q индуцирует у адресата, в частности, ожидание, что ситуация типа Q влияет на положение дел так, как это всегда бывает. Иными словами, ожидание соответствует знанию, как ситуация типа Q влияет на положение дел (например, как температура воздуха влияет на выбор человеком одежды). Это может быть ожидание, что имеет место 'не-P' (например, когда воздух становится теплым, человек не носит теплую одежду). Вторая пропозиция (P) может соответствовать этому знанию, и тогда высказывание оформляется соединительным союзом, ср. (1а). Но пропозиция P может противоречить этому знанию, и это противоречие маркируется противительным союзом *но* или *а*, ср. (1б). Аналогичные примеры:

(3) *Шли сильные дожди* (Q)*, и дороги размыло* (P).

(4) *Шли сильные дожди* (Q)*, но/а дороги не размыло* (P).

(5) *Был сильный мороз* (Q)*, и Ваня мерз* (P).

(6) *Был сильный мороз* (Q)*, но/а Ваня не мерз* (P).

Знания удобно записывать в виде фреймов или сценариев. С некоторым упрощением, знание действительности — это знание фреймов (сценариев). В разбираемых случаях знание — это общеизвестные представления о каузальных связях нашего мира, например, «в теплую погоду люди носят более легкую одежду, чем в холодную», «мороз действует на человека так-то и так-то», «сильные дожди размывают дороги». Такие каузальные связи в норме не нарушаются — они относятся к законам природы. Ожидание, индуцированное знанием законов природы, можно назвать объективным.

В подобных высказываниях соединительный и противительный союзы заменимы с большим трудом (напомним, что сейчас мы рассматриваем сложносочиненное предложение как отдельное сообщение, т. е. вне контекста). Ср.

(1a) ??*Стало тепло* (Q)*, но я снял шубу* (P).

(2a) ??*Стало тепло* (Q)*, и я не снимал шубу* (P).

(3a) **Шли сильные дожди* (Q)*, но/а дороги размыло* (P).

(4a) **Шли сильные дожди* (Q)*, и дороги не размыло* (P).

(5a) **Был сильный мороз* (Q)*, но Ваня мерз* (P).

(6a) **Был сильный мороз* (Q)*, и Ваня не мерз* (P).

Однако парадоксальным образом в некоторых случаях союзы *и* и *но* (или *а*) взаимозаменимы. Перейдем к этому материалу.

2.2. Ожидание и субъективная норма

Тот факт, что в каких-то высказываниях союзы *и* и *но /а* взаимозаменимы, отмечается В. З. Санниковым (1989). Ср.

(7) *Коля пошел на охоту* (Q)*, но вернулся с пустыми руками* (Q).

(8) *Коля пошел на охоту* (Q) *и вернулся с пустыми руками* (Q).

Возможная интерпретация этого факта такова: некоторые нормы, или «законы жизни» являются более жесткими, ср. материал из п. 2.1, а некоторые — менее жесткими. Нарушение менее жесткой нормы маркируется (противительным союзом) необязательно [Санников 1989]. Приведем аналогичные примеры:

(9) *Катя купила юбку* (Q)*, но/и она ей не подошла* (P).

(10) *Он хотел уехать из Москвы* (Q)*, но /и не уехал* (P).

(11) *Она сказала, что вышла замуж* (Q)*, но/и это была неправда* (P).

Требуется понять, в чем состоит содержательное различие между этими двумя типами норм.

На наш взгляд, в случаях типа (7)–(11) ситуация типа Q, хотя и влияет на положение дел, но сама по себе не обуславливает ситуацию типа не-P. Так, ситуация 'субъект купил что-либо' сама по себе не влечет за собой ситуацию 'у субъекта появилась нужная хорошая вещь'. Но такое положение дел представляется нам правильным, естественным, поэтому ожидание такого положения дел можно назвать «идеальным», или субъективным. Оно базируется на некоей идеальной картине мира. В этой картине мира зафиксированы не законы природы, а представления об идеальном, «образцовом» положении дел в человеческом обществе. Ясно, что такое ожидание вполне может нарушаться и нарушается. Его нарушение маркируется противительным союзом, однако он легко заменим на соединительный, ср. (8), а также (9)–(11). При этом соответствие ожиданию маркируется только соединительным союзом, ср.

(12) *Катя купила юбку* (Q)*, и/*но она ей подошла* (P).

Субъективное ожидание — это необязательно ожидание хорошего: в идеальной картине мире определенные ситуации влекут за собой плохое. Приведем пример (из книги В. З. Санникова (1989)):

(13) *Он заболел* (Q)*, и/но выздоровел* (P).

(14) *Он заболел* (Q)*, и/* но умер* (P).

Ситуация 'субъект заболел' индуцирует субъективное ожидание плохого: субъект будет болеть долго, если не всегда, или даже умрет. Поэтому нормальны высказывания: *Он заболел и умер* [соответствие ожиданию] и *Он заболел, но выздоровел* [несоответствие ожиданию]. Однако это ожидание, как и всякое субъективное ожидание, вполне может нарушаться. Поэтому нормальны высказывания: *Он заболел и выздоровел*, а также *Он тяжело болел и не умер*. Однако плохо: **Он заболел, но умер*, потому что описываемое положение дел соответствует идеальной картине мира (болезнь влечет за собой плохие последствия, возможно смерть), и это не может маркироваться союзом *но*.

На первый взгляд, ожидание в (13)–(14) кажется странным — оно расходится с нашими бытовыми представлениями, на что обратил внимание В. З. Санников (1989). Однако семантическая система языка (или, выражаясь иначе, языковая картина мира) сложилась достаточно давно и не всегда поспевает за изменениями жизни (в частности за успехами в медицине). В языковой

картине мира (по крайней мере, русской) болезнь по-прежнему ассоциируется не с выздоровлением, а с длительным продолжением или даже со смертью.

Для описания выбора союза *и* vs *но/а* нам понадобилось ввести представление о двух нормах: объективной и идеальной. Однако расщепление понятия «норма» на два более частных понятия необходимо и для описания других, причем больших групп лексики. Это, прежде всего, само слово *норма* и его аналоги: *образец, стандарт, эталон, шаблон*. Это также большой класс параметрических и оценочных прилагательных; ср. *большой, маленький, широкий, узкий, высокий, низкий* и т. п., *хороший, плохой, верный, удобный, вкусный* т. п. (Урысон 2011б). Таким образом, предлагаемое описание употребления союзов *и* vs *но/а* системно: оно базируется на понятиях, используемых при описании других классов языковых единиц. Заметим, что понятия объективной и идеальной нормы соответствуют двум интепретациям нормы в философии — философское понятие объективной (статистической) нормы разработано Спинозой, а философское представление об идеальной норме — Юмом (Арутюнова 1988).

Естественно предположить, что для адекватного представления языковых фактов требуется предусмотреть в базе знаний два «раздела»: в одном отражается объективная картина мира, а в другом — идеальная, или субъективная картина мира (представление об идеале). В объективной картине мира зафиксированы физические, химические, биологические и другие естественно-научные законы — ненарушаемые законы природы. В субъективной картине мира фиксируются представления о нормальном, т. е. хорошем социуме, о человеческих действиях и их нормальных следствиях, о нормальных человеческих взаимоотношениях, а также о человеческой жизни, ее протекании и т. п. Действительность часто не соответствует этим представлениям, т. е. субъективная картина мира отражает не законы, а всего лишь наши, часто не вполне осознаваемые ожидания относительно поведения людей, о жизни и смерти и т. п.

Можно предположить, что оба раздела базы знаний содержит утверждения вида: Если P, то Q. Например: Если идут сильные дожди, то дороги размывает; Если погода теплая, люди носят легкую одежду; и т. п. Однако субъективная картина мира описывается утверждениями, снабженными знаком вроде «По субъективному человеческому представлению». Например: По субъективному человеческому представлению, если кто-то покупает себе новую вещь, она ему подходит; По субъективному человеческому представлению, если кто-то заболевает, он может болеть долго или умереть. При этом сценарии, предполагаемые объективной картиной мира, ненарушаемы (их нарушение люди обычно считают чудом), а сценарии субъективной картины мира, напротив, скорее двойственны (ожидается, чтобы было так, но бывает и иначе).

Различия между сочинением, основанным на «объективной норме», и сочинением на основе «идеальной нормы» в некоторой степени отражается в перифразах сложносочиненных предложений.

Сложносочиненное предложение, указывающее на соответствие объективной норме, легко перифразируется в сложноподчиненное (благодарим за это замечание анонимного рецензента). Ср.

(3) *Шли сильные дожди, и дороги размыло — Дороги размыло, потому что шли сильные дожди.*

Кроме того, такие предложения допускают вставку слов или выражений *поэтому, в результате, вследствие этого* и т. п. непосредственно после союза *и*. Ср. *Шли сильные дожди, и поэтому <и вследствие этого, и в результате> дороги размыло.*

Что касается сложносочиненных предложений, отсылающих к идеальной норме, то многие из них подобных перифраз не допускают. Ср. *Катя купила юбку, и она ей (не) подошла — *Юбка (не) подошла Кате, потому что она ее купила*; * *Катя купила юбку, и поэтому она ей (не) подошла*. Однако для многих подобных предложений такие перифразы возможны. Ср.

(15) *Они прилетели в Филадельфию* (Q)*, и Лиза с ними там повидалась* (P).

В соответствии с предлагаемой интерпретацией, пропозиция Q индуцирует здесь субъективное ожидание ситуации P. Действительно, при нарушении такого ожидания в сложносочиненном предложении возможны как противительный союз *но*, так и соединительный союз *и*.

Ср. (16) *Они прилетели в Филадельфию (Q), но/и Лиза там с ними не повидалась (P)*.

Между тем, (15) допускает практически те же перифразы, что и (3). Ср.

(17) *Они прилетели в Филадельфию* (Q)*, и Лиза там с ними повидалась* (P). — *Лиза с ними повидалась, потому что они прилетели в Филадельфию. — Они прилетели в Филадельфию, и поэтому Лиза с ними там повидалась.*

Дело в том, что пропозиции в (15) связаны «слабым отношением обусловленности» по (ван Дейк 1978), благодаря чему этот пример и допускает перифразы, характерные для предложений, отражающих объективную картину мира. В связи с этим возникает задача выяснить, допускает ли объективная картина мира, т. е. законы природы, отношение слабой обусловленности. Если нет — то это отношение является принадлежностью лишь субъективной картины мира, и тогда оказывается, что соответствующие предложения в той или иной степени могут сближаться со сложносочиненными предложениями, отражающими объективную картину мира.

Найти четкие и тем более формальные различия между двумя описываемыми классами сложносочиненных предложений пока не удалось.

3. Ожидание и законы организации текста

Первая пропозиция Q сложносочиненного предложения активизирует у адресата не только некоторый фрагмент знания действительности, но и знание законов текста, в частности, законов развития повествования. Это знание индуцирует у адресата ожидание того, что (а) продолжение будет «на ту же

тему» или (б) в фокусе сохранится тот же объект. Подобное ожидание обусловлено тем, что, восприняв Q, адресат настраивается на данную тему/на данный объект. Эта настроенность обусловлена спецификой нашего восприятия информации, некоторой инерционностью нашего сознания. Поэтому такое ожидание можно условно назвать «инерционным». Оно сближается с хорошо известными в психологии установками восприятия (Узнадзе 1949). Пример:

(18) *В лесу теперь тоскливо* (Q)*, и дома тоже скучно, заняться нечем* (P).

Обе пропозиции здесь — на одну тему («скучно»), поэтому данное предложение оформляется соединительным союзом. Данное требование к пропозициям, соединяемым английским and ‘и’, впервые было отмечено в работе (Lakof R. 1971). Подробнее о понятии темы (микротекста) см. Урысон 2011.

Однако в первой пропозиции в фокусе внимания — лес, а во второй — дом. Смена фокуса маркируется противительным союзом; в современном русском языке это союз *а* (Kalkova, Podlesskaya 2001; Урысон 2002; 2004; 2011а). Поэтому в (18) возможна замена союза *и* на союз *а*. Ср.

(19) *В лесу теперь тоскливо* (Q)*, а дома тоже скучно, заняться нечем* (P).

Что именно маркировать (сохранение темы или смену объекта в фокусе), выбирает говорящий — это его стратегия построения сложносочиненного предложения.

Инерционное ожидание не нарушается даже при несоответствии P сценарному или субъективному ожиданию, т. к. вторая пропозиция P сохраняет некоторую общую микротему («что происходило в описываемое время»). Однако сценарное ожидание сильнее инерционного — оно больше влияет на выбор союза. Поэтому нарушение сценарного ожидания в отдельном сообщении маркируется обязательно, т. е. в этом случае всегда выбирается противительный союз (хотя теоретически могло бы маркироваться сохранение инерционного, т. е. выбираться соединительный союз); ср. (1а), (2а), (3)–(5).

Напротив, в случае нарушения субъективного ожидания говорящий может выбирать, маркировать ли нарушение этого ожидания или сохранение инерционного, ср. (7) и (8), а также (9)–(11). Иными словами, в случаях типа *Катя купила юбку, и она ей не подошла* союз *и* маркирует сохранение темы микротекста («что имело место в описываемое время»).

Языки могут различаться стратегиями, предпочитаемыми в тех или иных случаях. Так, в определенных контекстах в русском языке почти обязательно маркировать (союзом *а*) смену объекта в фокусе. Ср.

(20) *Что делают дети? — Ваня читает, а Маша рисует.*

В английском в подобных случаях обычно употребляется and, т. е. нормально маркируется общность микротемы пропозиций («занятие детей в данный момент»). Ср.

(21) *Kate is drawing, and Peter is playing with the cat.*

Подробнее о таких стратегиях (см. Rudnitskaya, Uryson 2008).

4. Сложносочиненное предложение и контекст

Если сложносочиненное предложение является частью текста, то на выбор союза влияет также предтекст (одна из первых работ на эту тему — статья ван Дейк 1978; наши данные отличаются от разбираемых ван Дейком). Предтекст может индуцировать свое ожидание, и оно важнее сценарного. Ср.

(22) *Мы условились, что когда станет тепло, я не буду снимать шубу. Стало тепло* (Q)*, и/*но я не снимал шубу* (P) [субъективное ожидание: договор не нарушают].

В тексте важна связность, поэтому даже при нарушении сценарного ожидания говорящий может маркировать сохранение инерционного. Ср.

(23) *Утром начался дождь* (Q)*, но/а детей повели на пляж* (P) [сценарное ожидание: детей не повели на пляж; несоответствие этому ожиданию]

vs.

(24) *Утром начался дождь* (Q)*, и детей повели на пляж* (P). *Многие впервые увидели море в непогоду* [соответствие инерционному ожиданию].

Предложение является частью текста, поэтому описание предложения как отдельного сообщения, вне какого-либо контекста, — это дань грамматической традиции. Такое описание может служить лишь подспорьем для описания предложения как фрагмента текста.

5. Заключение

Мы попытались объяснить, почему в некоторых сложносочиненных предложениях возможна замена противительного союза *но/а* на соединительный союз *и*. Для этого нам пришлось ввести понятие о двух типах норм — объективной норме, отражающей законы природы, и субъективной норме, отражающей наши представления о том, как должны действовать люди, как должна протекать человеческая жизнь и т. п. Соответствие норме всегда маркируется союзом *и*. Что касается нарушения нормы, то оказалось, что оно маркируется не обязательно.

Всегда маркируется (союзом *но/а*) несоответствие объективной норме. Ср. *Стоял сильный мороз, и река замерзла* [соответствие норме] — *Стоял сильный мороз, но река не замерзла* [несоответствие норме]. Между тем нарушение

субъективной нормы может маркироваться союзом (*но/а*), но может вообще не отмечаться. Ср. *Она сказала, что вышла замуж, и это была правда* [соответствие норме] — *Она сказала, что вышла замуж, но/и это была неправда* [несоответствие норме].

Найти другие надежные формальные различия между выделенными двумя типами сложносочиненных предложений не удалось. Однако предложенная интерпретация материала согласуется с описанием некоторых других, достаточно больших и разнородных фрагментов лексики, для которых также оказалось необходимым расщепить понятие нормы на два более узких (Урысон 2011б). Тем самым, предлагаемое описание выбранных союзов системно: оно использует понятия, необходимые для представления целого ряда классов слов.

Естественно предположить, что для адекватного представления языковых фактов требуется предусмотреть в базе знаний два «раздела»: в одном отражается объективная картина мира, а в другом — субъективная картина мира (представление об идеале).

Норма тесно связана с ожиданием: знанием нормы индуцируется определенное ожидание того, что имеет место или будет иметь место. Казалось бы, понятие ожидания избыточно для описания выбранных союзов. Однако выяснилось, что это не так: существует большой класс сложносочиненных предложений с союзами *и, но/а*, которые естественно описывать, используя понятие ожидания, а не нормы.

Дело в том, что в целом классе случаев ожидание вызвано не представлением о норме, а обусловлено некоторой инерционностью нашего сознания, его настроенностью на определенное продолжение высказывания. Это согласуется с общеизвестной теорией установок восприятия (Узнадзе 1948). Сохранение такого ожидания тоже маркируется союзом *и*, а его нарушение союзом *но* или *а*. Ср. *Петя готовится к экзаменам, и Вася заболел (в футбол поиграть не с кем)* [союз *и* маркирует общую тему «во дворе нет знакомых ребят»] vs. *Петя готовится к экзаменам, а Вася заболел (вот почему не с кем в футбол поиграть)* [союз *а* маркирует смену объекта в фокусе: ‘Петя’ vs. ‘Вася’]. Выбор союза в данном случае обусловлен стратегией построения высказывания, которая отчасти предопределена конкретным языком, но отчасти выбирается говорящим.

Нам удалось описать достаточно разнородный класс сложносочиненных предложений, используя единое понятие ‘ожидание’. Само ожидание может быть разной природы, т. е. может быть обусловлено разными факторами. Природа ожидания определяется исходя из энциклопедического значения конъюнктов. Таким образом, наше описание отличается от описания В. З. Санникова (1989), в котором в качестве исходного понятия используется понятие нормы. Мы подтвердили мысль Е. В. Падучевой (2004), что данный компонент является строевым компонентом семантической системы языка (или системообразующим смыслом по Апресяну). Преимущество предложенного подхода перед описаниями отдельных сочинительных союзов в том, что все они описываются с помощью единого базового понятия.

Для описания выбора союза *и* vs. *но/а* оказалось удобным не «портретировать» данные союзы, выделяя в них разные значения, а работать с каждым союзом, в том числе и с союзом *и*, как с цельной единицей. Такой подход может оказаться особенно удобным для описания художественных, в частности стихотворных, текстов.

Предложенное описание является теоретической моделью, которую необходимо верифицировать с помощью психологичессикх и, возможно, других экспериментов.

Работа поддержана грантами РГНФ № 13-04-00307а и № 15-04-00441а, НШ-3899.2014.6, а также грантом ОИФН РАН.

Литература

1. *Апресян Ю. Д.* 1974. Лексическая семантика. М.: «Наука».
2. *Арутюнова* 1988 — Н. Д. Арутюнова. Типы языковых значений: Оценка. Событие. Факт. М.: «Наука».
3. *Вольф Е. М.* 1986. Функциональная семантика оценки. М.: «Наука».
4. *Дейк ван* 1978. Вопросы прагматики текста // НЗЛ, вып. VIII. Лингвистика текста. М.: «Прогресс» [van Dijk, Issues in the pragmatics of disciurse. Univ. of Amsterdam, 1975].
5. *Левин Ю. И.* 1970. Об одной группе союзов русского языка // Машинный перевод и прикладная лингвистика. Вып. 13. М.: МГПИИЯ им. М. Тореза.
6. *Падучева Е. В.* 2004. Динамические модели в семантике лексики. М.: ЯСК.
7. *Санников В. З.* 1989. Русские сочинительные конструкции. М.: «Наука».
8. *Сепир Э.* 1944/1985. Градуирование: семантическое исследование // НЗЛ. Вып. XIV. М. : «Прогресс» [Sapir. Grading, a study in semantics // Philosophy of science. Vol. 11. 1944. № 2.]
9. *Узнадзе Д. Н.* 1949. Экспериментальные основы психологии. Тбилиси.
10. *Урысон Е. В.* Русский союз и частица И: структура значения // ВЯ. 2000. № 3.
11. *Урысон Е. В.* Союз А как сигнал поворота повествования // Логический анализ языка. Семантика начала и конца/Отв. ред. чл.-корр. РАН Н. Д. Арутюнова. М. 2002.
12. *Урысон Е. В.* Некоторые значения союза А в свете современной семантической теории // Русский язык в научном освещении. 2004. № 2.
13. *Урысон Е. В.* Семантика союза НО: данные языка о деятельности сознания // Вопросы языкознания. 2006, № 5.
14. *Урысон Е. В.* 2011а. Опыт описания семантики союзов. М.: ЯСК.
15. *Урысон Е. В.* 2011б. НОРМА, ОБРАЗЕЦ, ЭТАЛОН, СТАНДАРТ, ШАБЛОН: заметки о полисемии // Слово и язык. Сборник статей к восьмидесятилетию академика Ю. Д. Апресяна. М.: ЯСК. 2011.
16. *Blakemore D.* 1987. Semantic Constraints on Relevance. Oxford: Blackwell.
17. *Carlson L.* 1985. Dialogue games. An approach to discourse analysis. Dordrecht: Reidel.

18. *Kalkova T., V. Podlesskaya* 2001. The order of syntactic constituents vs. the order of discourse units: the case of Russian adversative constructions // Item order: its variety and linguistic and phonetic consequences. Prague: Karolinum Press.
19. *Kitis, E.* 2000. Connectives and frame theory // Pragmatics and cognition, vol. 8, N 2.
20. *Lakof R.* 1971. If's, and's, and but's about conjunction // Studies in linguistic semantics. N.Y: Holt, Reinhart, Williams.
21. *Rudnitskaya E., E. Uryson* 2008. Toward a semantic typology of coordination // Subordination and coordination strategies in North Asian languages/Ed. E. J. Vajda. John Benjamins Publ. Company: Amsterdam — Philadelphia.

References

1. *Apresyan Yu. D.* (1974), Lexical semantics [Leksicheskaya semantika], Nauka, Moscow.
2. *Arutyunova N. D.* (1988), Types of language meanings: Evaluation. Event. Fact. [Tipy yazykovykh znacheniy: Otsenka. Sobytiye. Fakt], Nauka, Moscow.
3. *Blakemore D.* (1987), Semantic Constraints on Relevance. Oxford: Blackwell.
4. *Carlson L.* (1985), Dialogue games. An approach to discourse analysis. Dordrecht: Reidel.
5. *Dijk van* (1975), Issues in the pragmatics of discourse. Univ. of Amsterdam.
6. *Kalkova T., V. Podlesskaya* (2001), The order of syntactic constituents vs. the order of discourse units: the case of Russian adversative constructions, in Item Order: its Variety and Linguistic and Phonetic Consequences. Prague: Karolinum Press.
7. *Kitis, E.* (2000), Connectives and frame theory, Pragmatics and cognition, vol. 8, N 2.
8. *Lakof R.* (1971), If's, and's, and but's about conjunction, Studies in linguistic semantics. N.Y: Holt, Reinhart, Williams.
9. *Levin Yu. I.* (1970), On a group of Russian conjunctions [Ob odnoy gruppe soyuzov russkogo yazyka], Machine translation and applied linguistics [Mashinnyy perevod i prikladnaya lingvistika], Vol. 13, Moscow.
10. *Paducheva E. V.* (2004), Dynamic models in lexical semantics [Dinamicheskiye modeli v semantike leksiki], Yazyki slavyansrikh kul'tur, Moscow.
11. *Rudnitskaya E., Uryson E.* (2008). Toward a semantic typology of coordination, in Subordination and Coordination Strategies in North Asian Languages, Ed. E. J. Vajda, John Benjamins Publ. Company: Amsterdam — Philadelphia.
12. *Sannikov V. Z.* (1989), Russian coordinating constructions [Russkiye sochinitel'nye konstruktsii], Nauka, Moscow.
13. *Sapir E.* (1944), Grading, a study in semantics, Philosophy of science, Vol. 11, № 2.
14. *Uryson E. V.* (2000), The Russian conjunction and particle I 'and': structure of meaning [Russkiy soyuz I chastitsa I: struktura znacheniya], Questions of linguistics [Voprosy yazykoznaniya], № 3.
15. *Uryson E. V.* (2002), The Russian conjunction A as a marker of a narration turn [Soyuz A kak signal povorota povestvovaniya], Logic analysis of language: Semantics of the beginning and of the end [Logicheskiy analiz yazyka: semantika nachala i kontsa], Indrik, Moscow.

16. *Uryson E. V.* (2004), Some meanings of the Russian conjunction A in the light of the modern semantic theory [Nekotorye znacheniya soyuza A v svete sovremennoy semanticheskoy teorii], Russian language in scientific elucidation [Russkiy yazyk v nauchnom osveshchenii], № 2.
17. *Uryson E. V.* (2006), The meaning of the Russian conjunction NO 'but': language data on mind functioning [Semantika soyuza NO: dannye yazyka o deyatel'nosti soznaniya], Questions of linguistics [Voprosy yazykoznaniya], № 5.
18. *Uryson E. V.* (2011a), A trial of description of conjunctions meaning [Opyt opisaniya semantiki soyuzov], Yazyki slavyansrikh kul'tur, Moscow.
19. *Uryson E. V.* (2011б), The Russian words NORMA 'norm', OBRAZETS 'example', ETALON 'sample', STANDART standard", SHABLON 'pattern': notes on polysemy, in Word and Language, Yazyki slavyansrikh kul'tur, Moscow.
20. *Uznadze D. N.* (1949), The experimental foundations of psycology [Experimental'nye osnovy psikhologii], Metsniereba, Tbilisi.
21. *Vol'f E. M.* (1986), Functional semantics of evaluation [Funktsional'naya semantika otsenki], Nauka, Moscow.

ТЕЗАУРУСЫ РУССКОГО ЯЗЫКА В ВИДЕ ОТКРЫТЫХ СВЯЗАННЫХ ДАННЫХ

Усталов Д. А. (dmitry.ustalov@urfu.ru)

Уральский федеральный университет имени первого Президента России Б. Н. Ельцина, Екатеринбург, Россия;
NLPub, Екатеринбург, Россия

Важной тенденцией последних лет являются открытые лингвистические данные, дающие исследователям и разработчикам возможность построения собственных решений на основе готовых и выверенных словарей, корпусов, тезаурусов, и других ресурсов. При этом опубликованные данные хранятся в разных форматах, что затрудняет их эффективное использование, а также привязывает пользователей к поставщику. Данная работа посвящена представлению популярных тезаурусов русского языка в виде открытых связанных данных: описаны существующие форматы данных и подходы к их преобразованию, выполнено отображение трёх популярных открытых русских тезаурусов в схемы Семантической паутины. Полученный набор данных опубликован в формате Turtle и доступен на ресурсе NLPub для использования на условиях лицензии Creative Commons.

Ключевые слова: связанные данные, открытые данные, лексические ресурсы, интеграция данных, семантическая паутина, русский язык

RUSSIAN THESAURI AS LINKED OPEN DATA

Ustalov D. A. (dmitry.ustalov@urfu.ru)

Ural Federal University, Yekaterinburg, Russia;
NLPub, Yekaterinburg, Russia

Open linguistic data is a good recently established trend allowing both researchers and developers in the field of natural language processing to create their own applications using high-quality dictionaries, thesauri, corpora, etc. At the same time, the published open data are stored in different formats making them difficult to be used in an efficient way without falling within vendor lock-in. This paper is devoted to the problem of representing popular lexical resources of the Russian language in the form of Linked Open Data. It summarizes the recent work in the field of thesauri representation formats and approaches to converting such formats to those of Linked Data. It also proposes an approach to converting popular Russian thesauri to the vocabularies that are the essential parts of the Linguistic Linked Open Data Cloud. The proposed approach has been implemented in open source software and the resulted dataset has been made publicly available on NLPub in the Turtle format under the terms of a Creative Commons license.

Key words: Linked Data, Open Data, lexical resources, data integration, Semantic Web, Russian language

1. Introduction

Open linguistic data is a good recently established trend allowing both researchers and developers in the field of natural language processing to create their own applications using high-quality dictionaries, thesauri, corpora, etc. At that, the published open data are stored in different formats making them difficult to be used in an efficient way without falling within vendor lock-in. Hence, both the Semantic Web and natural language processing for Russian fields could benefit from representing the popular Russian thesauri in the form of Linked Data allowing applications to use the Semantic Web technologies including the powerful reasoning tools.

The work, as described in this paper, makes the following contributions: 1) it summarizes the recent work in the fields of thesauri formats, thesauri conversion approaches, and the thesauri for Russian, 2) it proposes and implements an approach to convert popular Russian thesauri to the form Linked Data, and 3) presents the results under a Creative Commons license. The rest of this paper is organized as follows. Section 2 is devoted to the survey on the related work. Section 3 proposes an approach to converting the thesauri to the Linked Data representation. Section 4 describes the implementation, presents and evaluates the resulted dataset. Section 5 concludes with final remarks and directions for the future work.

2. Related Work

The following three directions of the related work are considered: 1) thesauri representation formats, 2) approaches for converting thesauri into Linked Data, and 3) publicly available electronic thesauri for Russian.

2.1. Representation Formats

Princeton WordNet, the most recognized and influential electronic lexical ontology, has been represented in the form of semi-structured text files [3]. This format is widely used and the majority of WordNet's derivatives[1] utilize it to keep compatibility with the original database. In spite of the WordNet's popularity, there are many software libraries for various programming languages allowing one both to read and write the linguistic data in this format. However, dealing with such a format has a significant drawback: embedding it into an application requires either conversion into the form of relational database or utilizing special software to integrate the present data schema with WordNet's. This makes the resulting data model less uniform and requiring additional maintenance. There also exist many linguistic resources operating with their in-house developed custom data formats, hence their formats will be denoted as *custom*.

XML, eXtensible Markup Language, is designed to be both human-readable and machine-readable [12]. It is used by many production-grade software in order

[1] http://globalwordnet.org/

to describe data in almost every existing domain including lexical resources. The main advantages of XML are its wide support, representation uniformity, and the ability to be validated against a predefined schema. However, processing of large XML documents containing hundreds megabytes of data is computationally hard as it requires either construction of an expensive document object model (DOM) to be stored entirely in a computer memory, or through the use of stream-oriented SAX parsers, which are much less convenient in development.

Resource Description Framework (RDF) is proposed to be a uniform representation of any subject-predicate-object entity for the Semantic Web [13]. The RDF is just an abstract syntax that should be encapsulated by serialization formats, e. g. RDF/XML, Turtle, N3, etc. It should be noted that the Simple Knowledge Organization System (SKOS) is an RDF extension designed especially for vocabularies and thesauri [11]. The syntax of RDF triplets is simple to be understood by human, but it is difficult for an end user to support all the available representation formats. RDF/XML is the most popular hence it has the same drawbacks as XML.

The recently published ISO 25964 standard is designed to formalize fitting concepts, terms and relationships together to make a thesaurus [7]. It is focused on the knowledge engineering aspect of thesauri and does not propose a representation format leaving such a task to the user. Another standard, ISO 24613:2008 describes LMF, a lexical markup framework, which is aimed at providing an XML-based representation for vocabularies without dealing with word senses and their relations [4].

2.2. Linked Data Conversion

The problem of converting a thesaurus into the Linked Data has been approached for several times since the appropriate data schemas have been issued.

van Assem et al. in 2006 proposed a method to convert a thesaurus to the SKOS format and assessed the applicability of such a representation [10]. The proposed method has three steps: 1) analyzing thesaurus, 2) mapping data items into SKOS, and 3) creating a conversion program. This method has been evaluated on three thesauri (IPSV, GTAA and MeSH) and it has been confirmed that SKOS is suitable for thesauri resembling to the ISO 25964 standard.

McCrae et al. in 2011 presented a model called lemon (Lexical Model for Ontologies) that supports sharing terminological and lexicon resources on the Semantic Web [6]. The lemon model is an RDF-native form making it possible to expose a thesaurus to the Linked Data in the similar way as LMF does, and also represents word senses and their relations.

In 2012, Navigli & Ponzetto released BabelNet, which is a very large multilingual semantic network constructed automatically on the basis of WordNet, Wikipedia and other databases [8]. BabelNet integrates seamlessly into the Semantic Web[2] through the alignment to underlying data sources, and exposes itself in the form of RDF employing such vocabularies as SKOS and lemon.

[2] http://babelnet.org/rdf/

2.3. Thesauri for Russian

There are three notable electronic thesauri for Russian that are publicly available under open licenses: 1) RuThes-lite, 2) the Russian Wiktionary, 3) the Universal Dictionary of Concepts, and 4) Yet Another RussNet (the more detailed survey is presented on [5, 9]).

RuThes-lite[3] is a subset of the RuThes lexical ontology having been developed since 1994 for addressing the information retrieval tasks in various applications for the Russian language [5]. The format of the original RuThes is unknown, nevertheless RuThes-lite is available under the terms of the CC BY-*NC*-SA license in the form of quasi-structured HTML pages on the Internet representing approximately 26,000 concepts and 100,000 relations between them.

The Universal Networking Language[4] is a project led by the United Nations dedicated to the development of a computer language that replicates the functions of natural languages. The Russian version of its semantic network—the Universal Dictionary of Concepts—is contributed by the researchers from IITP RAS [2]. UNLDC is distributed under the CC BY-SA license containing approximately 62,000 of the universal words (UWs) and 90,000 links between them.

The Russian Wiktionary[5] is the eighth largest Wiktionary composed of more than 520,000 articles—one article represents a lexical entry—written by more than 120,000 users (only 164 users are active participants) since 2004. The native format of the Wiktionary pages is a quasi-structured wiki syntax, which is quite hard to parse. However, there exists the Wikokit[6] project that parses the Russian and English Wiktionaries and renders them in the machine-readable form of a relational database available under the terms of the CC BY-SA license.

Yet Another RussNet[7] is an open project established in 2013 and aimed at creation of a large electronic thesaurus for Russian through the use of crowdsourcing [1]. At the moment, this resource contains 111,895 words and approximately 18,000 synsets. Initially, the project deliverables had been available in the XML format with the correspondent XSD although recently the synsets have been made available in the CSV and RDF formats, which are more convenient to parse and to use. All the content is published under the CC BY-SA license. Yet Another RussNet includes the lexicon and synsets of the Russian Wiktionary among several others resources licensed under the same license.

It should be noted that all these resources utilize their own *custom* data representation formats embarrassing their evaluation and forcing end users into vendor lock-in. Since that the Yet Another RussNet project includes the lexicon and synsets of the Russian Wiktionary, only three resources will be considered in this study: RuThes-lite, UNLDC, Yet Another RussNet.

---

[3] http://www.labinform.ru/pub/ruthes/index.htm

[4] http://www.undl.org/

[5] https://ru.wiktionary.org/

[6] https://code.google.com/p/wikokit/

[7] http://russianword.net/

3. Representing the Thesauri

In order to represent the above-mentioned resources in the form of Linked Data, it is necessary to make the following assumptions. Firstly, the primary applications of the present work are natural language processing and information retrieval, thus the resulted resource may not cover the complete set of natural language entities and relations. Secondly, the resulted dataset should not reinvent the Linked Data vocabularies, but should use the popular ones as soon as possible. Finally, both humans and machines should easily understand the resulted data format.

Each thesaurus has been analyzed to find out how the data items can be mapped to the Linked Data vocabularies, and each thesaurus will be presented in a separate ontology. Since the RuThes-lite thesaurus is widely applied in various practical tasks, the types of its concept relations have been considered as the only concept relation types with one exception: the antonymy relation, which is widely used in UNLDC.

The choice of the Linked Data vocabularies is mostly inspired by that of BabelNet, hence the following vocabularies have been used: Simple Knowledge Organization System (SKOS) to represent concepts, Lexicon Model for Ontologies (lemon[8]) to represent lexical senses, lexical entries, definitions and usage examples, LexInfo[9] to represent the morpho-syntactic labels. RDFS, OWL and Dublin Core have expressed the ontology description. The Turtle format has been chosen to store the processing output because of its readability and popularity.

Table 1 demonstrates the result of thesaurus entities' mapping. The most challenging part of the mapping process was the selection of the appropriate concept relation representation. For instance, SKOS provides the special terms for expressing hypernymy and hyponymy, but does not provide such terms for holonymy and meronymy—although LexInfo does.

Table 1. Entities, relations, vocabularies

Entity/Relation	Vocabulary Term
Concept	*skos*:Concept
Lexical Sense	*lemon*:LexicalSense
Definition (Gloss)	*lemon*:SenseDefinition
Usage Example	*lemon*:UsageExample
Lexical Entry	*lemon*:LexicalEntry
Lemma	*lemon*:Form
Class-Subclass (is-a)	*skos*:{broader,narrower}
Part-Whole (part-of)	*lexinfo*:{holonymTerm,meronymTerm}
Asymmetric Association	*lemon*:subsense
Symmetric Association	*skos*:related
Antonomy	*lexinfo*:antonym
Sense-Concept Mapping	*lemon*:{sense,isSenseOf}

[8] http://www.lemon-model.net/

[9] http://lexinfo.net/

Entity/Relation	Vocabulary Term
Sense-Lexeme Mapping	*lemon*:{reference,isReferenceOf}
Sense-Definition Mapping	*lemon*:definition
Sense-Example Mapping	*lemon*:example
Lemma Indication	*lemon*:canonicalForm
Part-of-Speech	*lexinfo*:partOfSpeech

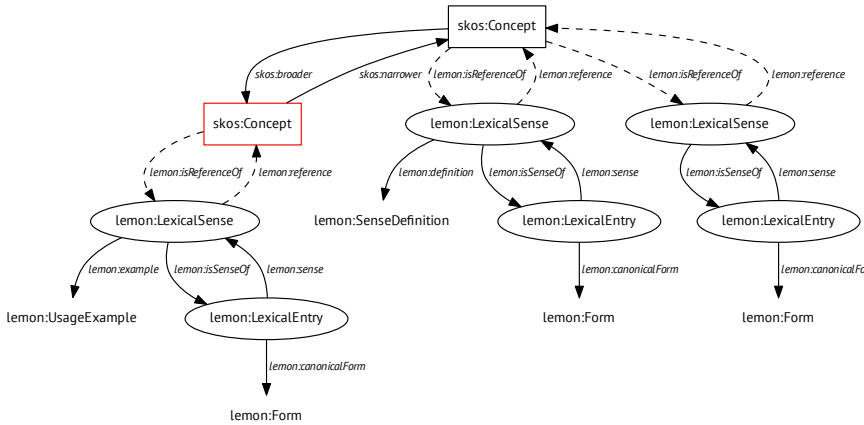


Fig. 1. Elements of the uniform ontology: concepts, lexical entries and their senses, definitions, usage examples, and lemmas

An example of the resulted ontology is depicted at Fig. 1. The present example shows two defined concepts: one has two lexical senses and is denoted as a hypernym to the other (red colored) concept having only one sense. Each sense is provided with the corresponding lexical entries. Each lexical entry hopefully has a canonical form (lemma). The picture also illustrates that the sense definitions and the usage examples are connected to the lexical senses instead of the concepts.

3.1. RuThes-lite

RuThes-lite comes in the form of four schema-less XML files representing lexical entries, concepts, their relations, and mappings between the concepts and lexemes. Despite RuThes-lite containing some valuable morpho-syntactic information, it is written in barely parsable form, and such information has been—unfortunately—omitted, i.e. the fields like `synt_type` and `pos_string`. Table 2 summarizes the mapping process.

Table 2. Mapping RuThes-lite to the Linked Data vocabularies

Data Item	Feature/Function	Property/Class
//entry[@id]	Lexical Entry Lemma Lemma Indication	*lemon*:LexicalEntry *lemon*:Form *lemon*:canonicalForm
//concept	Concept	*skos*:Concept
//entry_rel	Lexical Sense Sense-Concept Mapping Sense-Lexeme Mapping	*lemon*:LexicalSense *lemon*:{sense,isSenseOf} *lemon*:{reference,isReferenceOf}
//rel[@name="ВЫШЕ"]	Class-Subclass (is-a)	*skos*:broader
//rel[@name="НИЖЕ"]		*skos*:narrower
//rel[@name="ЧАСТЬ"]	Part-Whole (part-of)	*lexinfo*:holonymTerm
//rel[@name="ЦЕЛОЕ"]		*lexinfo*:meronymTerm
//rel[@name="АСЦ1"]	Asymmetric Association	*lemon*:subsense
//rel[@name="АСЦ2"]		*lemon*:subsense
//rel[@name="АСЦ"]	Symmetric Association	*skos*:related

3.2. The Universal Dictionary of Concepts

UNLDC is published[10] in the form of CSV files representing universal words (UWs) and links between them. Since the UNLDC universal words are unambiguous by design, they have been mapped into lexical senses as described in Table 3. The main problem of the UNLDC mapping is the necessity to parse the domain-specific relations stored within such UWs as *tongue(icl>concrete_thing,pof>body)*, therefore such descriptors were omitted and the resulted dataset has no relations. The synsets[11] derived from the UWs have been mapped to concepts.

Table 3. Mapping UNLDC to the Linked Data vocabularies

Data Item	Feature/Function	Property/Class
Lemma	Lexical Entry Lemma Lemma Indication	*lemon*:LexicalEntry *lemon*:Form *lemon*:canonicalForm
Part-of-Speech	Part-of-Speech	*lexinfo*:partOfSpeech
Universal Word	Concept Lexical Sense Sense-Concept Mapping Sense-Lexeme Mapping	*skos*:Concept *lemon*:LexicalSense *lemon*:{sense,isSenseOf} *lemon*:{reference,isReferenceOf}
Relation icl	Class-Subclass (is-a)	*skos*:{broader,narrower}
Relation iof		
Relation pof	Part-Whole (part-of)	*lexinfo*:{holonymTerm,meronymTerm}
Relation equ	Symmetric Association	*skos*:related
Relation ant	Antonomy	*lexinfo*:antonym

[10] https://github.com/dikonov/Universal-Dictionary-of-Concepts/tree/master/data/csv

[11] https://github.com/dikonov/Universal-Dictionary-of-Concepts/tree/master/data/misc

3.3. Yet Another RussNet

The Yet Another RussNet software is implemented in the Ruby on Rails framework with active use of the ActiveRecord object-relational mapping [9]. Table 4 shows that its data models[12] have been mapped to those of Linked Data.

Table 4. Mapping Yet Another RussNet to the Linked Data vocabularies

Data Item	Feature/Function	Property/Class
Word	Lexical Entry Lemma Lemma Indication Part-of-Speech	*lemon*:LexicalEntry *lemon*:Form *lemon*:canonicalForm *lexinfo*:partOfSpeech
Synset	Concept	*skos*:Concept
SynsetWord	Lexical Sense Sense-Concept Mapping Sense-Lexeme Mapping Sense-Definition Mapping Sense-Example Mapping	*lemon*:LexicalSense *lemon*:{sense,isSenseOf} *lemon*:{reference,isReferenceOf} *lemon*:definition *lemon*:example
Definition	Definition (Gloss)	*lemon*:SenseDefinition
Example	Usage Example	*lemon*:UsageExample

4. Results

The conversion and the supplementary programs have been implemented in the Ruby programming language. The resulted software is available on GitHub under the MIT license: https://github.com/nlpub/rtlod. During the implementation, it has become necessary to port the lemon and LexInfo vocabularies to the syntax of the used RDF.rb library, which resulted in releasing of the *rdf-lemon*[13] library for Ruby.

The resulted dataset consisting of the converted RuThes-lite, UNLDC and Yet Another RussNet thesauri in the Turtle format is available on NLPub: http://nlpub.ru/RTLOD. Thorough evaluation of such a resource is a very interesting topic that is complicated enough to conduct a specialized study. Nevertheless, in order to compare the resulted ontologies quantitatively, brief statistics of them have been calculated and demonstrated in the Table 5. The lexical intersection between the converted thesauri has also been assessed (Table 6).

It seems that Yet Another RussNet that is created through crowdsourcing has the widest lexical coverage although the number of its concepts is relatively low. It is also the only resource provided with the word usage examples due to its crowdsourcing schema requiring users to consider such examples [1]. High number of lexical senses is caused by the presence of many duplicated synsets generated by users. Despite this resource having no established concept relations, it still may be still useful

[12] http://nlpub.ru/YARN/API

[13] https://github.com/nlpub/ruby-rdf-lemon

as a synonyms' dictionary in some applications. Both RuThes-lite and UNLDC are mature resources with developed concept relations [5], but UNLDC is a dictionary of a controlled language [2], hence its number of concepts is significantly smaller, although these concepts are tightly connected to each other.

Table 5. Resulted datasets

# of	RuThes-lite	UNLDC	Yet Another RussNet
Lexical Entries	96,700	56,313	111,895
Part-of-Speech Tags	n/a	56,313	111,821
Concepts	26,354	8,896	17,492
Relations	98,976	n/a	0
Lexical Senses	115,106	20,366	69,981
Definitions	10,701	8,896	7,641
Usage Examples	0	0	2,991

Table 6. Lexical intersection

		# of common lexical entries
RuThes-lite	UNLDC	18,596
UNLDC	Yet Another RussNet	26,088
Yet Another RussNet	RuThes-lite	37,920

5. Conclusion

The author believes that the present work—especially the published dataset and software—could facilitate the development of the modern linguistic resources for Russian among their integration into the Linguistic Linked Open Data Cloud[14]. Given the openly published resources, a user can choose between them in order to pick the best option for the particular application. Moreover, it contributes a lot into simplifying conducting thorough studies of these resources by such benchmarks as word sense disambiguation competitions. The present mapping approach is general and could be freely used for adopting more thesauri of the Russian language.

There are several reasons for future work. Firstly, it may be useful to assess the lexical coverage of the given resources with these representations. Secondly, since these datasets are Linked Data, it may be interesting to estimate alignments between concepts of them. Finally, end users may consume the deliverables of this work and link their own data to these.

[14] http://linghub.lider-project.eu/llod-cloud

Acknowledgements

This work is supported by the Russian Foundation for the Humanities, project № 13-04-12020 "New Open Electronic Thesaurus for Russian". The author would like to thank Natalia V. Loukachevitch from the Lomonosov Moscow State University for the provided RuThes-lite dataset in the XML format, and Vladimir V. Ivanov from the Kazan Federal University and Vyacheslav Dikonov from the Institute for Information Transmission Problems of the RAS for fruitful discussions and valuable suggestions. The author is also grateful to the anonymous referees who offered very useful comments on the present paper.

References

1. *Braslavski P., Ustalov D., Mukhin M.* (2014), A Spinning Wheel for YARN: User Interface for a Crowdsourced Thesaurus, Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics, Gothenburg, Sweden, pp. 101–104.
2. *Dikonov V. G.* (2013), Development of lexical basis for the Universal Dictionary of UNL Concepts, Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue", Bekasovo, Russia, pp. 212–221.
3. *Fellbaum C.* (1998), WordNet: An Electronic Lexical Database, MIT Press, Cambridge, MA, USA.
4. *Francopoulo G.* (2013), LMF: Lexical Markup Framework, Wiley-ISTE, London, UK.
5. *Loukachevitch N. V., Dobrov B. V., Chetviorkin I. I.* (2014), RuThes-Lite, a Publicly Available Version of Thesaurus of Russian Language RuThes, Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue", Bekasovo, Russia, pp. 340–349.
6. *McCrae J., Spohr D., Cimiano P.* (2011), Linking Lexical Resources and Ontologies on the Semantic Web with Lemon, Springer: Lecture Notes in Computer Science, Vol. 6643, pp. 245–259.
7. *National Information Standards Organization*. (2011), ISO 25964 – the international standard for thesauri and interoperability with other vocabularies, available at: http://www.niso.org/schemas/iso25964/
8. *Navigli R., Ponzetto S. P.* (2012), BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network, Artificial Intelligence, Vol. 193, pp. 217–250.
9. *Ustalov D.* (2014), Enhancing Russian Wordnets Using the Force of the Crowd, Springer: Communications in Computer and Information Science, Vol. 436, pp. 257–264.
10. *van Assem M., Malaisé V., Miles A., Schreiber G.* (2006), A Method to Convert Thesauri to SKOS, Springer: Lecture Notes in Computer Science, Vol. 4011, pp. 95–109.
11. *World Wide Web Consortium W3C*. (2004), SKOS Simple Knowledge Organization System, available at: http://www.w3.org/2004/02/skos/
12. *World Wide Web Consortium W3C*. (2008), Extensible Markup Language (XML) 1.0 (Fifth Edition), available at: http://www.w3.org/TR/2008/REC-xml-20081126/
13. *World Wide Web Consortium W3C*. (2014), RDF 1.1 Concepts and Abstract Syntax, available at: http://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/

СТАТЬЯ ЗНАЧИТ СТАТЬЯ: ОБ ОДНОМ КЛАССЕ ТАВТОЛОГИЧЕСКИХ КОНСТРУКЦИЙ В РУССКОМ ЯЗЫКЕ [1, 2]

Вилинбахова Е. Л. (elenavilinb@yandex.ru)

Санкт-Петербургский государственный университет, Санкт-Петербург, Россия

Ключевые слова: тавтологии, метаязыковые конструкции, русский язык, интернет

ARTICLE MEANS ARTICLE: ON ONE PATTERN OF TAUTOLOGIES IN RUSSIAN

Vilinbakhova E. L. (elenavilinb@yandex.ru)

St. Petersburg State University, St. Petersburg, Russia

The study based on Internet and corpus data [RNC] deals with a special pattern of tautological constructions in Russian called metalinguistic tautologies. The notion was briefly introduced in 1996 by E. Miki as a label for a set of quite heterogeneous examples, but was not further developed. While other tautologies describe entities in the real world, metalinguistic tautologies refer to the use of a linguistic expression. Such constructions show that the speaker is employing a word or an expression in its common, straight meaning. Therefore, they are most often used when context allows other possible interpretations of the linguistic expressions (such as euphemisation, irony, or hyperbole), and sometimes such alternatives are explicitly spelled out: *Inexpensive means inexpensive, not poor quality.* Metalinguistic tautologies are established in Russian with patterns *X znachit X*, *X oznachaet X* 'X means X', *X eto X* 'X is X', and are distinguished from homonymous constructions by their semantic and pragmatic features.

Key words: tautologies, metalinguistic constructions, Russian language, Internet

---

[1] Автор выражает признательность анонимным рецензентам статьи за их замечания и предложения.

[2] Работа поддержана грантом Президента РФ для государственной поддержки ведущих научных школ РФ «Школа общего языкознания Ю. С. Маслова» НШ-1778.2014.6.

Если дипломат говорит «да», это значит «может быть»,
если дипломат говорит «может быть», это значит «нет»,
если дипломат говорит «нет», значит, это не дипломат.

Если женщина говорит «нет», это значит «может быть»,
Если женщина говорит «может быть», это значит «да»,
если женщина говорит «да», значит, это не женщина.

Если военный говорит «да», это значит «да»,
если он говорит «нет», это значит «нет»,
если военный говорит «может быть», значит, это не военный.

(Анекдот)

1. Вводные замечания

Тавтологические конструкции вида *Жизнь есть жизнь* неоднократно становились объектом лингвистических исследований. Особенно активно обсуждались вопросы, связанные с механизмом их интерпретации, возможностью их перевода с одного языка на другой без потери смысла, влиянием контекста и лексического наполнения конструкций на возможные инференции и т.д., см. [Grice 1975], [Levinson 1983], [Wierzbicka 1987], [Gibbs, McCarrell 1990], [Ward, Hirschberg 1991], [Miki 1996], [Bulhof, Gimbel 2001], [Падучева 2004], [Meibauer 2008], [Rhodes 2009] и др.

На материале русского языка Т. В. Булыгина и А. Д. Шмелёв выделяют два ряда тавтологических конструкций: *X (и) есть X, X всегда / везде X* и т. п., с одной стороны, и *X — это X*, с другой стороны: «значение конструкций первого ряда: все манифестации X-а, т.е. члены класса X или «инстанты» индивида X, в общем, одинаковы, и нет оснований ждать от манифестации, с которой мы имеем дело, чего-то другого. <…> Для *X — это X* общее значение состоит в «выделении X-а среди прочих индивидов или классов; констатация его особого места; особенности отдельных манифестаций могут не приниматься во внимание» [Булыгина, Шмелёв 1997: 506–509].

Можно заметить, что Т. В. Булыгина и А. Д. Шмелёв связывают значения обоих рядов тавтологий с определёнными синтаксическими моделями, однако первичными являются именно семантические признаки: идея равноценности членов внутри класса для первого ряда и идея исключительности класса для второго ряда. Отметим, что конструкции обоих рядов используются говорящим для выражения мнения по поводу объектов или явлений окружающей действительности.

Тавтологии, о которых пойдёт речь, образуют третий ряд и могут быть проиллюстрированы примерами (1) и (2). В настоящей работе они обозначаются как *метаязыковые тавтологии* (далее — МТ), термин заимствован из [Miki 1996].

(1) *А: Leo, тут надо разобраться, что в вашем понимании есть рыбалка!!!*
Б: ***Рыбалка значит рыбалка****…. РЫБУ ЛОВИТЬ…..и народ ухой угостить…*[3]

[3] Примеры взяты из Интернета, если не указано иное. Из соображений краткости адреса не указаны, но, разумеется, в материалах они присутствуют.

(2) *А: Что значит «завтра»? Он должен взять билет.*
*Б: **Завтра означает завтра**. Успеет купить.*

Статья построена следующим образом: после краткой истории вопроса (раздел 2) будут описаны особенности МТ, отличающие их от других тавтологических конструкций (раздел 3); разновидности МТ и их общее значение (раздел 4); репрезентации МТ в русском языке и формально сходные конструкции, которые МТ не являются (раздел 5). В разделе 6 представлены итоги исследования.

2. История вопроса

Первое упоминание о МТ встречается в статье [Miki 1996], посвящённой значению тавтологических конструкций. Э. Мики исследует, каким образом «бессмысленные» тавтологические высказывания обретают значение в коммуникации, и приходит к выводу, что в тавтологиях набор известных собеседникам характеристик объекта идентифицируется в контексте текущей ситуации с самим объектом. Автор указывает на существование особых случаев, которые отсылают к языковому коду участников коммуникации, и обозначает их как «метаязыковые» тавтологии (англ. *metalinguistic tautologies*). В статье даны следующие примеры:

«(33) *When I said kid, I meant K-I-D, kid! 'Когда я сказал ребёнок, я имел в виду Р-Е-Б-Ё-Н-О-К, ребёнок'.*

(35) *When he said three o'clock, he meant three o'clock* (P. G. Wodehouse, The Romance of an Ugly Policeman) *'Когда он сказал «три часа», он имел в виду три часа'.*

(37) *Partial means partial 'Неполный значит неполный'.*

(41) *I'm glad to meet you, I said, meaning the words.* (J. Braine. Room at the top) *'Очень рад познакомиться с вами, — сказал я, и это были не пустые слова: я действительно был рад'* (пер. Т. Озерской и Т. Кудрявцевой).

(42) *I don't have one. / — What do you mean you don't have one? / — I mean I don't have one* (C. Webb. The Graduate) *'У меня его нет / Что значит, у тебя его нет?/ Это значит, у меня его нет'*

(43) *We are getting sick and tired of it. When I say he's in Chicago, he is in Chicago.* (F. S. Fitzgerald. The Great Gatsby) *'Покою нет от вашего брата. Раз я говорю, он в Чикаго, значит, он в Чикаго'* (пер. Е. Калашниковой)»

[Там же: 641].

Можно заметить, что приведённые Э. Мики примеры разнородны по форме и особенно по содержанию: ср. (35) и (43). В первом случае говорящий указывает,

что языковое выражение *три часа* в устах героя действительно имело буквальное значение ‘три часа’, а не ‘2.50’ или ‘пол четвёртого’, т. е. он имел в виду ровно то, что сказал, и этот частный смысл совпадает с общепринятым значением выражения *три часа*. Во втором случае говорящий подчёркивает достоверность предоставляемой слушателю информации и убеждает посетителя в том, что босса *действительно* нет на месте, т. е. в данном контексте языковое выражение *он в Чикаго* является истинным. Отметим, что для случаев, аналогичных примеру (35), соответствие содержания языкового выражения и реальной действительности не обязательно:

(35’) ***Когда он сказал «три часа», он,*** *конечно,* ***имел в виду три часа****, но, к сожалению, из-за пробок смог добраться только к половине четвёртого.*

(43’) ****Если я говорю, он в Чикаго, значит, он в Чикаго****, но на самом деле он Калифорнии.*

По-видимому, эти противоречия можно было бы устранить при более тщательной разработке понятия МТ, однако идея не получила дальнейшего развития ни у самого автора, ни у его коллег: класс МТ не упоминается в обзорных частях дальнейших исследований, посвящённых тавтологиям, см. [Bulhof, Gimbel 2001, [Meibauer 2008], [Rhodes 2009].

На русском материале вопрос о МТ был затронут в [Вилинбахова 2013], однако с тех пор удалось сделать некоторое количество дополнений и уточнений.

3. Метаязыковые тавтологии в ряду других тавтологических конструкций

Может показаться, что в определённых контекстах МТ очень близки по смыслу к конструкциям первого и второго ряда, описанным в [Булыгина, Шмелёв 1997]: ср. конструкцию первого ряда (3) и МТ (4) с общей импликатурой ‘необходимость выполнения обязательств’.

(3) *Но* ***обещание — есть обещание****, слово следует держать.*

(4) *А: Что значит — обещание?*
Б: ***Обещание значит обещание****, то, что нарушить нельзя.*

Тем не менее, МТ имеют свои особенности, выделяющие их в ряду остальных тавтологий. Наиболее очевидное отличие — это связка ‘обозначать, иметь значение’, с помощью которой в МТ даётся характеристика языковому знаку, а не объектам действительности, как в тавтологиях первого и второго ряда. Кроме того, отличия касаются языковых выражений, используемых в качестве повторяющихся элементов МТ: это их автонимный, а не термовый,

референциальный статус[4] и их употребление в первом, «наименее переносном» значении.

Как было указано выше, тавтологии первого и второго ряда описывают ***внеязыковую действительность***. Рассмотрим следующие примеры: в (5) говорящий признаёт, что семья занимает важное место в жизни человека, и в частности — его собеседника, который пропускает дружеские встречи; в (6) подразумевается, что семья — это надёжный тыл, т. е. в обоих высказываний речь идёт о семье как о социальном явлении.

(5) ***Семья есть семья,*** *но общаться надо и с друзьями, выкидывать нельзя людей из жизни!*

(6) *Даже если не получиться, ты всегда успеешь вернуться домой, ведь* ***семья — это семья****.*

МТ, напротив, дают инструкцию к восприятию ***языкового выражения***: см. (7), где речь идёт о слове *семья*, толкование которого говорящий даёт в следующей реплике.

(7) *А: Что для Вас значит семья?*
Б: ***Семья — значит семья****! Семья — это дорогие люди, которые всегда будут рядом, что бы ни случилось в жизни.*

Таким образом, в тавтологиях, описанных Т. В. Булыгиной и А. Д. Шмелёвым, первый из повторяющихся элементов имеет термовый референциальный статус, а в МТ — автонимный референциальный статус[5].

Различия в референциальном статусе элементов определяют формальные ограничения на языковые выражения, допустимые в тавтологических конструкциях. Если в тавтологиях первого и второго ряда преобладают именные группы и невозможны, например, союзы или частицы, которые не относятся к референциальным выражениями, то МТ не имеют таких ограничений, см. (8). Также в МТ в качестве повторяющихся элементов могут выступать предикативные конструкции, см. (9) и (10).

---

[4] Референциальный статус именных групп, вводящих в рассмотрение ту или иную внеязыковую сущность, называется **термовым**. <…> Референциальный статус имени, которое отсылает не к той или иной сущности в мире и не к свойству, а к самому себе, называется **автонимным** [Кобозева 2009: 229].

[5] С этой точки зрения МТ напоминают «нетавтологические» примеры (см. ниже), где для читателей расшифровываются потенциально незнакомые слова. Однако, в отличие от реальных толкований, МТ не дают новых данных, а отсылают к уже известной собеседнику информации.
Это слово происходит от итальянского «дилетто», означающее «удовольствие» <…>. ***Дилетант — значит неспециалист****, точнее, не получивший специального образования в той отрасли науки, где он отваживается что-то сказать. [НКРЯ]*

(8) *А: Ой. / Б: Что ой? / А:* ***Ой значит ой*** *и глаза закрой.*

(9) *Ну у нас всё просто…* ***хочу есть значит хочу есть****… а не своди меня в ресторан)))*

(10) *Что значит «не люблю»? — Экий ты, брат, непонятливый, — ответил за Лену Майоров, делая еще один шаг в их сторону. —* ***«Не люблю» значит «не люблю»****, ни больше, ни меньше.*

Следующая особенность МТ связана с вопросом о характере общих фоновых знаний собеседников и нормативности употребления языковой единицы. Неоднократно отмечалось, что при использовании тавтологических конструкций говорящий даёт отсылку к некоторой информации, хорошо знакомой слушателю, см. [Gibbs, McCarrell 1990], [Miki 1996], [Bulhof, Gimbel 2001], [Meibauer 2008]. Это могут быть общекультурные знания, принятые в языковом коллективе, убеждения, распространённые в кругу отдельной семьи, друзей или коллег, и, наконец, факты, которые только что обсуждались участниками коммуникации. Для тавтологий первого и второго ряда речь идёт о свойствах объекта, поэтому подразумеваемая информация могут быть очень «закрытой» (знать о конкретных свойствах объекта могут только собеседники), а нормативность употребления языкового выражения не имеет значения, ср. (11), где существительное *вышка* используется в значении ‘высшая лига’.

(11) ***Вышка это вышка****, не нужно путать её с Украинской лигой.*

В случае с МТ идёт отсылка к языковой компетенции адресата, которая также входит в фонд общих знаний собеседников, но уже не может быть «локальной», т. к. это противоречит семантике конструкции. Дело в том, что говорящий, используя МТ, представляет значение языкового выражения как общеизвестное, очевидное, простейшее, поэтому языковые выражения должны не просто присутствовать в сознании носителей языка, но и использоваться в своём первом, «наименее переносном» значении.

Устойчивый, общепринятый характер значения языковых единиц может быть подчёркнут апелляцией к словарю как «единому источнику нормативных сведений о литературном языке, который не пополняется новыми данными и, вообще говоря, содержит информацию, которая заведомо известна говорящему как достаточно образованному человеку» [Иомдин 2014: 130], см. пример (12) из данной работы:

(12) ***Новый это значит новый.*** (*Открой словарь русского языка*). *Если есть какие то сомнения то не покупай дешевле всё равно не найдёшь* [Иомдин 2014: 119]

В примере (13) у глагола *спать* имеются оба рассматриваемых значения, но второе является переносным и потому отметается говорящим.

(13) *ВНИМАНИЕ! Это реальное предложение, а не прикол. И, да, когда я говорю «спать» — это не завуалированный намек на интим-услуги.* ***Спать — означает спать.***

Поскольку в значение МТ входит апелляция к наиболее общепринятому значению, изучение данных конструкций могло бы пролить свет на то, какие значения многозначных слов говорящие считают главными, исходными, подобно тому как тавтологии первого и второго ряда используются для изучения стереотипных представлений об объектах и ситуациях, см., например, [Кобозева 2009: 185–197].

4. Разновидности метаязыковых тавтологий

Как уже отмечалось выше, МТ указывают на соответствие частного употребления языкового выражения общепринятому значению. Такое речевое поведение имеет смысл, когда возможна альтернативная интерпретация, см. тип апелляций к словарю в интернет-дискуссиях и его толкование в уже упомянутой работе Б. Л. Иомдина: «Говорящий употребляет слово X в общепринятом значении, но предполагает, что адресат может неправильно понять значение слова X, и на этом основании заранее оценивает поведение адресата отрицательно» [Иомдин 2014: 118].

Ср. также рассуждения Г. Уорда и Д. Хиршберг, которые считают, что наличие альтернативных вариантов является важнейшим условием употребления в речи любых, а не только метаязыковых, тавтологий: «используя в речи тавтологии, говорящий нарушает постулат количества (информативности), позволяя слушателю сделать вывод, что намеренно не были выбраны определённые альтернативные высказывания. <…> Конкретные варианты отвергнутых альтернативных высказываний зависят от контекста» [Ward, Hirschberg 1990: 510].

В случае с МТ контексты, допускающие альтернативные варианты понимания языкового выражения, делятся, как минимум, на три группы, и, таким образом, можно говорить о трёх разновидностях конструкции.

4.1. МТ, отрицающие эвфемизацию

‘Говорящий подчеркивает, что произнесенная им языковая единица употреблена в наиболее общепринятом значении и не является эвфемизмом, как неправильно предполагает адресат’. Сюда относятся контексты, где языковая единица могла бы заменять собой другое, социально менее приемлемое выражение, как в (13) и (14), где *спать* и *кофе* означает вполне определённый досуг. В эту же группу попадают МТ, отрицающие различного рода намёки, см. (9).

(14) [обсуждение на тему: «Если мужчина приглашает на чашечку кофе»]
А с чего ты взяла, что секс вообще будет? ***Кофе — значит кофе****.*

4.2. МТ, отрицающие иронию

'Говорящий подчеркивает, что произнесенная им языковая единица употреблена в наиболее общепринятом значении, потому что адресат может заподозрить иронию и понять его неправильно'. В данных контекстах возможно шутливое, ироническое и пр. употребление языкового выражения, меняющее его смысл на противоположное, как в (14), где собеседник понимает слово *подарок* как 'товар за деньги'.

(15) — *Ладно, а я тебе, шотландо-фриз, между прочим, подарочек приготовил. — Он нагнулся и достал из-под прилавка толстенный том. — Кто-то из университетских профессоров заказал — и не выкупил. Словарь британских фамилий. Я думал, какому придурку мне толкнуть эту дурацкую книгу? И вспомнил про тебя.*
— Спасибо, <…>… Сколько сдерешь за подарок?
*— **Подарок — значит, подарок**. Наслаждайся, пока я добрый!*
Фабель поблагодарил Отто еще раз, уже не насмешливо…

4.3. МТ, отрицающие гиперболу

'Говорящий подчеркивает, что произнесенная им языковая единица употреблена в наиболее общепринятом значении, потому что адресат может предположить преувеличение / преуменьшение и т. п. и понять его неправильно'. Речь идёт о контекстах, где языковое выражение в силу различных обстоятельств может использоваться неточно, с допустимой долей погрешности, как в (15), где автор предполагает, что читатели могут понять выражение *чуть-чуть (согнуть)* как, скажем, 'до некоторой степени', а не 'совсем немного'.

(16) [Всероссийский виндсерфинг форум] *И чтобы доска хорошо ребрилась заднюю ногу действительно нужно чуть-чуть сгибать. Но **чуть-чуть значит «чуть-чуть»**, т. е. только чтобы загнуть плавник.*

Можно заметить, что предложенные разновидности МТ соотносятся с коммуникативными постулатами Г. П. Грайса [Grice 1975]: говорящий настаивает, что нарушение, например, постулата истинности не имело место, поскольку, по его мнению, адресат может быть уверен в обратном. МТ указывают, что в данном контексте возможно другое, не буквальное восприятие языкового выражения, однако это не соответствует замыслу говорящего. Таким образом, можно описать общее значение МТ: 'Говорящий подчеркивает, что произнесенная им языковая единица употреблена в наиболее общепринятом значении и не содержит обычных отклонений от него'[6] (эти отклонения — эвфемизация, ирония и гипербола — и соответствуют трем приводимым разновидностям конструкций).

---

[6] Автор благодарен за эту формулировку анонимному рецензенту.

5. «Истинные» и «ложные» метаязыковые тавтологии

Как видно из примеров в [Miki 1996], на материале английского языка класс МТ представлен различными синтаксическими моделями. В русском языке распространены модели *X значит X*, *X означает X*, а также модель *X — это X*, см. (17).

(17) *Воспитанный в строгих канонах западной дипломатии, когда «да» — это «да», а «нет» — это «нет», он не понимал всех хитростей и тонкостей политики на Востоке* [НКРЯ]

Пунктуационное оформление МТ в текстах разнообразно и определяется автором сообщения: возможны кавычки или курсив для одного или обоих повторяющихся элементов, запятая, тире, многоточие до и после связки и т. д.

Наиболее «удобной» для исследователя является модель *X означает X*, поскольку полученные при запросах данной модели примеры в большинстве случаев оказываются МТ. Возможность замены связки (*значит / это*) в остальных конструкциях на связку *означает* без изменения смысла зачастую может служить тестом на принадлежность их к МТ. Примеры по модели *X — это X* могут оказаться как МТ, так и тавтологиями второго ряда, выражающими идею исключительности класса (см. раздел 1). При запросе модели *X значит X,* помимо МТ, выдаются примеры омонимичной конструкции *X значит X* со значением неконтролируемости выбора и импликатурой ‘ситуацию изменить невозможно, и переживать по этому поводу бессмысленно’, сходную с конструкцией *X так X*, описанную в [Копотев, Файнвейц 2007].

(20) ***Двойка значит двойка!*** *Это мелочи жизни).*

В указанных выше случаях «истинные» МТ отграничиваются от омонимичных конструкций достаточно легко: «ложные» МТ обладают другим значением, а языковые выражения в них имеют термовый, а не автонимный референциальный статус.

Более сложным являются случаи вроде *(Если / раз) Y говорит X, значит (,) X, Сказано X — значит X* и т. п., которые Э. Мики относит к МТ, см. известный рекламный текст, аналогичный приведённому выше примеру (43) из [Мики 1996].

(18) *Муж: Лично я всё решаю сам. Как скажу, так и будет.* ***Если я говорю на футбол, значит на футбол****.*
Жена: А не стоит ли нам поехать к маме?
Муж: ***Если я говорю к маме, значит, к маме****.*

В разделе 2 уже было отмечено, что для подобных примеров важно соответствие значения языкового выражения (в данном случае, *на футбол* и *к маме*) действительности, которое для остальных МТ не играет роли. Более того, языковое выражение является лишь одним из компонентов ситуации, которая, как показывает

маркер *значит*, «служит говорящему основанием для того, чтобы он сделал речевой акт логического вывода» [Храковский 1998 цит. по: Апресян и др. 2010: 220]. Строго говоря, намеренность общения вообще не является обязательным при порождении языкового выражения в данных конструкциях, см. (17), где тесты показывают определённый результат, который соответствует действительности.

(19) *Ненене, они* [тесты] *никогда не врут!* ***говорят добрый — значит добрый****.*
Я им теперь знаешь как верю.

Значение данных конструкций можно сформулировать следующим образом: ‘В случае, если языковое выражение X порождается в заданном контексте, можно сделать вывод, что имеет место ситуация, которая совпадает с планом содержания X’, что не соответствует рассмотренному выше значению МТ и, следовательно, позволяет не относить их к классу МТ.

6. Выводы

В русском языке, наряду с описанными Т. В. Булыгиной и А. Д. Шмелёвым двумя рядами тавтологических конструкций со значениями равноценности членов внутри класса и исключительности класса, представлен также класс метаязыковых тавтологий вида *Нет значит нет*. Используя МТ в речи, говорящий подчеркивает, что произнесенная им языковая единица употреблена в наиболее общепринятом значении и не содержит обычных отклонений от него, а именно: эвфемизации, иронии и гиперболы. В отличие от других тавтологий, языковые выражения в МТ имеют автонимный референциальный статус, что позволяет использовать в данных конструкциях любые языковые единицы. В русском языке МТ представлены синтаксическими моделями *X значит X*, *X означает X* и *X — это X*, которые отграничиваются от омонимичных конструкций по семантико-прагматическим признакам.

Литература

1. *Апресян Ю. Д., Богуславский И. М., Иомдин Л. Л., Санников В. З.* (2010), Теоретические проблемы русского синтаксиса: взаимодействие грамматики и словаря, Языки славянских культур, Москва.
2. *Булыгина Т. В., Шмелёв А. Д.* (1997), Языковая концептуализация мира (на материале русской грамматики), Школа «Языки русской культуры», Москва.
3. *Вилинбахова Е. Л.* (2013), Доклад значит доклад: метаязыковые тавтологические конструкции в русском языке, Тезисы конференции «Русский язык: конструкционные и лексико-семантические подходы», Санкт-Петербург, с. 9–10, режим доступа: www.iling.spb.ru/confs/rusconstr2013/pdf/abstracts.pdf
4. *Кобозева И. М.* (2009), Лингвистическая семантика, Книжный дом «ЛИБРОКОМ», Москва.

5. *Копотев М. В., Файнвейц А. В.* (2007), Изучать так изучать: синхрония и диахрония, Научно-техническая информация, Серия 2, Информационные процессы и системы, Вып. 9, с. 29–37.
6. *Иомдин Б. Л.* (2014), Загугли в Дале. Словари в интернет-дискуссиях, в Современный русский язык в интернете, «Языки славянской культуры», Москва, с. 114–132.
7. Национальный корпус русского языка, режим доступа: www.ruscorpora.ru
8. *Падучева Е. В.* (2004), Динамические модели в семантике лексики, «Языки славянской культуры», Москва.
9. *Bulhof J., Gimbel S.* (2001), Deep tautologies, Pragmatics and Cognition, Vol. 9–2, pp. 279–291.
10. *Gibbs R. W., McCarrell N. S.* (1990), Why boys will be boys and girls will be girls: understanding colloquial tautologies, Journal of Psycholinguistic Research, Vol. 19, pp. 125–145.
11. *Grice H. P.* (1975), Logic and conversation, in Syntax and Semantics, Vol. 3: Speech Acts, Academic Press, New York, pp. 41–58.
12. *Levinson S.* (1983), Pragmatics, Cambridge University Press, Cambridge.
13. *Meibauer J.* (2008), Tautology as presumptive meaning, Pragmatics and Cognition, Vol. 16, pp. 439–470.
14. *Miki E.* (1996), Evocation and tautologies, Journal of Pragmatics, Vol. 25, pp. 635–648.
15. *Rhodes R.* (2009), A Cross-linguistic comparison of tautological constructions with special focus on English, available at: www.linguistics.berkeley.edu/~russellrhodes/pdfs/taut_qp.pdf
16. *Ward, G. L., Hirschberg J.* (1991), A pragmatic analysis of tautological utterances, Journal of Pragmatics, Vol. 15, pp. 507–520.
17. *Wierzbicka A.* (1987), Boys will be boys: 'Radical semantics' vs. 'Radical pragmatics', Language, Vol. 63, pp. 95–114.

References

1. *Apresyan Ju. D., Boguslavskiy I. M., Iomdin L. L., Sannikov V. Z.* (2010), Theoretical issues in Russian syntax: interaction of grammar and dictionary [Teoreticheskie problemy russkogo sintaksisa: vzaimodeystvie grammatiki i slovarya], Jazyki slavyanskikh kul'tur, Moscow.
2. *Bulygina T. V., Shmelev A. D.* (1997), Linguistic conceptualization of the world (on the material of the Russian grammar) [Yazykovaya kontseptualizatsiya mira (na materiale russkoy grammatiki)], Shkola "Jazyki russkoy kul'tury", Moscow.
3. *Bulhof J., Gimbel S.* (2001), Deep tautologies, Pragmatics and Cognition, Vol. 9–2, pp. 279–291.
4. *Gibbs R. W., McCarrell N. S.* (1990),Why boys will be boys and girls will be girls: understanding colloquial tautologies, Journal of Psycholinguistic Research, Vol. 19, pp. 125–145.
5. *Grice H. P.* (1975), Logic and conversation, in Syntax and Semantics, Vol. 3: Speech Acts, Academic Press, New York, pp. 41–58.

6. *Iomdin B. L.* (2014), Google in Dal. Dictionaries in Internet discussions [Zagugli v Dale. Slovari v internet-diskussiyakh], in Modern Russian language in Internet [Sovremennyy russkiy yazyk v internete], "Jazyki slavyanskoy kul'tury", Moscow, pp. 114–132.
7. *Kobozeva I. M.* (2009), Linguistic semantics [Lingvisticheskaya semantika], Knizhnyy dom "LIBROKOM", Moscow.
8. *Kopotev M. V., Faynveyts A. V.* (2007) If we study we study: both synchrony and diachrony [Izuchat' tak izuchat': sinkhroniya i diakhroniya], Scientific and technical information. Series 2, Informational processes and systems [Nauchno-tekhnicheskaya informatsiya, Seriya 2, Informatsionnye protsessy i sistemy], Vol. 9, pp. 29–37.
9. *Levinson S.* (1983), Pragmatics, Cambridge University Press, Cambridge.
10. *Meibauer J.* (2008), Tautology as presumptive meaning, Pragmatics and Cognition, Vol. 16, pp. 439–470.
11. *Miki E.* (1996), Evocation and tautologies, Journal of Pragmatics, Vol. 25, pp. 635–648.
12. *Rhodes R.* (2009), A Cross-linguistic comparison of tautological constructions with special focus on English, available at: www.linguistics.berkeley.edu/~russellrhodes/pdfs/taut_qp.pdf
13. Russian National Corpus [Natsional'nyy korpus russkogo yazyka], available at: www.ruscorpora.ru
14. *Paducheva E. V.* (2004) *Dynamic models* in lexical *semantics* [Dinamicheskie modeli v semantike leksiki], Jazyki slavyanskoy kul'tury, Moscow.
15. *Vilinbakhova E. L.* (2013), Presentation means presentation: metalinguistic tautologies in Russian [Doklad znachit doklad: metayazykovye tavtologicheskie konstruktsii v russkom yazyke], Proceedings of the Conference "The Russian language: constructional and lexico-semantic approaches" [Materialy konferentsii "Russkiy yazyk: konstruktsionnye i leksiko-semanticheskie podkhody], St. Petersburg, available at: www.iling.spb.ru/confs/rusconstr2013/pdf/abstracts.pdf
16. *Ward, G. L., Hirschberg J.* (1991), A pragmatic analysis of tautological utterances, Journal of Pragmatics, Vol. 15, pp. 507–520.
17. *Wierzbicka A.* (1987), Boys will be boys: 'Radical semantics' vs. 'Radical pragmatics', Language, Vol. 63, № 1, pp. 95–114.

«ДАВАЙ РОНЯТЬ СЛОВА»: МЕТАФОРА КАУЗАЦИИ ПЕРЕМЕЩЕНИЯ ПО ВОЗДУХУ В СЕМАНТИЧЕСКОЙ ЗОНЕ ГЛАГОЛОВ РЕЧИ В РУССКОМ ЯЗЫКЕ В КОНТРАСТИВНОМ АСПЕКТЕ

Яковлева И. В. (irinadubrov@gmail.com)

Ульяновский государственный университет, Ульяновск, Россия

Данная статья посвящена исследованию метафоры каузации перемещения по воздуху в семантической зоне глаголов речи в русском языке в контрастивном аспекте. Подобный семантический переход характерен для глаголов бесконтактного перемещения, подразумевающих контроль только в начальной фазе действия (*бросать*). В русском языке особое место среди каузативов перемещения по воздуху, обнаруживающих способность к рассматриваемому семантическому переходу, занимают предикаты, в семантике которых заложено представление о некотором действии, осуществляемом рукой (*бросать*). Другая зона, включающая глаголы, содержащие представление о выталкивании объекта через некоторое отверстие, преимущественно ротовое (*изрыгать*), в русском языке служит источником метафорического переноса значительно реже и оказывается ограниченной явными негативными коннотациями. У глаголов, подразумевающих выделение определенной субстанции всей поверхностью (*излучать*), в русском языке не развивается способность к исследуемому метафорическому переходу, и представление о такой гипотетической донорской зоне мы получаем, рассматривая данные русского языка на фоне других языков. При осуществлении данного семантического перехода действуют ограничения на вид глагола, обусловленные сочетаемостью базовых глаголов *говорить* и *сказать*.

Ключевые слова: семантический сдвиг, метафора, метонимия, ребрендинг, глаголы каузации перемещения, глаголы речи, лексическая типология

THE METAPHOR OF CAUSED MOTION IN THE AIR IN THE TARGET DOMAIN OF SPEECH ACT VERBS IN RUSSIAN FROM A CONTRASTIVE PERSPECTIVE

Yakovleva I. V. (irinadubrov@gmail.com)

Ulyanovsk State University, Ulyanovsk, Russia

The study is devoted to the metaphor of caused motion in the air in the semantic field of speech act verbs in Russian from a contrastive perspective. This semantic shift is typical for verbs of contactless motion implying contact only in the initial phase like *brosat'* ('to throw'). In Russian verbs implying an action performed with hands demonstrate the outstanding capability of accomplishing the semantic shift under study. Another source domain including predicates implying an ejection of a certain object through a kind of hole, mainly a mouth, is not very productive in Russian and is restricted with some negative connotations. The verbs implying emission of a kind of substance with the whole surface do not normally accomplish this type of semantic shift in Russian and we distinguish this hypothetic source domain only when we study the Russian data against the typological background. The semantic shift under study is subject to the aspectual restrictions brought about by the co-occurrence of the basic speech act verbs *govorit'* ('to speak') and *skazat'* ('to say').

Key words: semantic shift, metaphor, metonymy, rebranding, verbs of caused motion, speech act verbs, lexical typology

1. Введение

Как известно, семантическая группа глаголов речи весьма неоднородна. К собственно глаголам речи, также имеющим центр и периферию, что подробно рассматривается в исследованиях [Зализняк Анна А. 2013: 152–155; Кобозева 1985], примыкают глаголы других семантических зон, обнаруживающие способность к семантическому переходу в зону глаголов речи, особое место среди которых занимают глаголы каузации перемещения по воздуху, такие как *бросить, обронить* и т. д. И если эмоции метафорически сближаются с жидкостями, и в их семантическую зону включаются глаголы перемещения в воде, то словам естественным образом приписывается способность перемещаться по воздуху. Об этом свидетельствует то, что в некоторых языках такая способность заключена уже в семантической структуре базового глагола со значением 'говорить'. Так в адыгейском языке глаголы *Іон* 'говорить' и *Іун* 'распространиться, разлететься (о вести, звуке)' возводятся к одному историческому корню [Шагиров 1977:159–160]. Данные русского языка будут рассматриваться

в контрастивном аспекте, поскольку «другие языки, если их рассматривать с точки зрения сочетаемостных возможностей лексики, служат тем зеркалом, которое высвечивает противопоставления, незаметные внутри лексической системы одного языка» [Рахилина, Лемменс 2003: 314]. В нашей работе мы хотели бы рассмотреть семантические переходы, позволяющие глаголам из донорской зоны каузации передвижения по воздуху проникать в реципиентную зону глаголов речи.

2. Семантический сдвиг в структуре каузативов перемещения по воздуху

При переходе в семантическую зону глаголов речи каузативы перемещения по воздуху могут соответствовать следующим значениям глагола *говорить* в классификации Анны А. Зализняк: 1.1 ‘сообщать’ и 2.1 ‘произносить осмысленный текст’ [Зализняк Анна А. 2013: 155]. При этом рассматриваемые глаголы приобретают способность встраиваться в типичную для глаголов речи конструкцию $X_{агенс}$ V $Y_{содержание}$ $Z_{адресат}$. Нужно отметить, что валентность адресата является факультативной. Валентность содержания при данных глаголах может заполняться прямым объектом, придаточным предложением с союзом *что*, а также прямой речью.

В целом, каузативы перемещения по воздуху можно подразделить на глаголы, в семантике которых заложено представление о:

1) каком-либо действии, придающем ускорение объекту и совершаемом преимущественно с помощью руки (*бросать*);
2) выделении объекта через некоторое единичное отверстие, преимущественно ротовое (*выплевывать, извергать*);
3) выделении вещества всей поверхностью (*излучать, выделять*).

Рассмотрим способность глаголов данных групп переходить в семантическую зону глаголов речи и встраиваться в характерные для глаголов этой зоны конструкции.

2.1. Глаголы каузации перемещения по воздуху с помощью руки

Каузативы перемещения, обусловленного определенным движением руки, можно подразделить на предикаты, обозначающие контролируемое либо неконтролируемое действие, с одной стороны, и допускающие пассивность или активность объекта, с другой стороны. Так глаголы *бросать* и *кидать* подразумевают намеренность, контролируемость действия со стороны агенса в начале, а глаголы *ронять, обронить, проронить* характеризуются утратой агенсом контроля над ситуацией. При этом все указанные глаголы относятся к группе глаголов бесконтактного перемещения в терминологии Г. И. Кустовой [Кустова 2004: 155], в то время как для глаголов контролируемого перемещения (*ставить, класть*), семантика которых подробно рассматривается

в работе [Кустова 2004: 122–155], семантический переход в зону глаголов речи не характерен. Объект перемещения во всех рассматриваемых случаях оказывается пассивным. В то же время глагол *отпустить* предполагает намеренную утрату контроля над объектом, который в свою очередь может обладать некоторой активностью. Рассмотрим поведение глаголов в каждой из групп.

2.1.1. Группа *бросить*

Когда глагол *бросить* проникает в семантическую зону глаголов речи, происходит изменение таксономического класса самого предиката и его аргументов. Если первоначально подразумевалось контролируемое движение руки, каузирующее перемещение некоторого материального объекта по воздуху, то при переходе данного глагола в семантическую зону глаголов речи происходит сдвиг значения, обусловленный изменением таксономического класса его аргументов: теперь в качестве прямого объекта выступает не какой-либо материальный объект, а произносимая человеком членораздельная речь, при этом целью такого речевого акта является доведение информации до сведения адресата. Валентность содержания при данном глаголе может выражаться всеми способами, доступными для каузативов перемещения: прямым объектом, придаточным предложением с союзом *что* и прямой речью. Относительно перемещения глагола *бросить* в семантическую зону глаголов речи мы можем утверждать, что имеем дело с метафорой.

Употребление глагола *бросить* в конструкции $X_{\text{агенс}}$ V $Y_{\text{содержание}}$ $Z_{\text{адресат}}$ либо просто подчеркивает стремительность действия с оттенком небрежности, как в примере (1), либо предполагает негативную, пренебрежительную реакцию агенса по отношению к адресату, предмету речи или ситуации в целом, как в примере (2):

(1) *Он мог, заглянув мимоходом в какой-нибудь отдел и услыхав там разговор на любую тему, бросить реплику, другую, затем завладеть всеобщим вниманием и превратить беседу в собственный монолог.* (Евгений Рубин. Пан или пропал. Жизнеописание. 1999–2000)

(2) *«Ну, значит, я была права, — торжественно заявила она своей соседке, а мне бросила: — Вы свободны».* (Людмила Гурченко. Аплодисменты. 1994–2003)

Что касается глагола *бросать*, нужно отметить, что его функционирование в семантической зоне глаголов речи несколько ограничено. Дело в том, что глаголы совершенного вида, обозначающие события, лучше встраиваются в характерные для глаголов речи конструкции и чаще вводят прямую речь, чем глаголы несовершенного вида, обозначающие процесс. Это может быть связано с различием в употреблении базовых глаголов данного семантического поля *говорить* и *сказать*. Глагол *сказать* часто рассматривается как видовой коррелят глагола *говорить*. При этом разница в сочетаемости данных глаголов проявляется и в тех конструкциях, в которых способны употребляться оба предиката. Глагол *сказать* в целом оказывается более частотным. Так в НКРЯ найдено 588 539 его

употреблений наряду с 486 466 употреблениями глагола *говорить*. При этом глагол *сказать* вводит прямую речь в приблизительно в 238 тыс. случаев (40,4 % от всех употреблений), а глагол *говорить* в 112 тыс. контекстов (23 % от всех употреблений). Такое различие обусловлено тем, что рассматриваемая структура чаще всего описывает уже свершившееся действие, а для глаголов совершенного вида «самый естественный контекст — прошедшее время» [Падучева 2004: 7]. Именно глагол *сказать*, являясь перфективным коррелятом глагола *говорить*, обозначает в акциональном отношении событие и оказывается «привязанным» к прошедшему времени, как показано в [Рахилина (ред.) 2010: 36].

С этой точки зрения интересно отметить, что конструкция с придаточным изъяснительным с союзом *что* оказывается в этом смысле значительно более нейтральной: глагол *сказать* вводит придаточное изъяснительное с союзом *что* в 62 тыс. случаев (11 % от всех употреблений данного глагола), а соответствующие употребления с глаголом *говорить* насчитывают 48 тыс. контекстов (9 % от всех употреблений). Дело в том, что, как показывает Е. Н. Никитина для видовой пары *писать — написать*, несовершенный вид делает акцент на содержании речи [Никитина 2015: 843], и, вероятно, подобной функцией концентрации на содержании, а не на обстоятельствах самого акта речи обладает и конструкция с придаточным изъяснительным, но этот вопрос требует дальнейшей проработки.

Возвращаясь к глаголам каузации перемещения, отметим, что по данным НКРЯ количество контекстов, в которых глагол *бросить* вводит прямую речь, составляет 858 вхождений (92 % от всех употреблений данного глагола в семантической зоне глаголов речи). Что касается глагола *бросать*, который также способен встраиваться во все конструкции, доступные для глаголов каузации перемещения в семантической зоне глаголов речи, в НКРЯ насчитывается 158 случаев употребления данного глагола в конструкции с прямой речью (80 % от всех употреблений данного глагола в семантической зоне глаголов речи). Остальные употребления глаголов *бросить* и *бросать* приходятся на конструкцию с придаточным изъяснительным с союзом *что* (по данным НКРЯ для глагола *бросить* это 18 контекстов, а для глагола *бросать* — 4 вхождения, т. е. 2 % от всех употреблений в рассматриваемой семантической зоне в обоих случаях) и конструкцию с прямым объектом (для глагола *бросить* это 61 контекст, т. е. 6 % от всех употреблений, и 36 вхождений для глагола *бросать*, т. е. 18 %).

Показательной представляется не только разница в проценте употребления глаголов совершенного и несовершенного вида в конструкции с прямой речью, но и непосредственно количественная разница. Именно то, что глаголы совершенного вида как бы «притягиваются» конструкцией с прямой речью, обусловливает значительно более высокую способность глаголов совершенного вида к переходу в семантическую зону глаголов речи. Так для глагола *бросить* общее количество употреблений в данной семантической зоне составляет 937 вхождений, в то время как для глагола *бросать* — всего 198.

Таким образом, видовые противопоставления в конструкциях с глаголами речи, намечающиеся при сопоставлении сочетаемости базовых глаголов *говорить* и *сказать*, становятся очевидными при исследовании семантических переходов в данной лексической зоне.

Способностью к переходу в семантическую зону глаголов речи обладают также глаголы *кинуть* и *кидать*, однако такие употребления немногочисленны.

Что касается дериватов рассматриваемых глаголов, нужно отметить, что в некоторой степени такой способностью обладают глагол *выбрасывать* и его видовой коррелят *выбросить*. Однако их употребление весьма ограничено. Так при глаголе *выбрасывать* валентность содержания может заполнять только существительное *слова*, причем таких употреблений очень немного:

(3) *Она выбрасывала слова быстро, на крике, погрузив пальцы в волосы и теребя их, словно желая вырвать* (Н. Н. Шпанов. Ученик чародея (1935–1950)

Глагол *выбросить* окказионально обнаруживает способность вводить прямую речь, НКРЯ содержит 2 таких употребления:

(4) *И только, может быть, через полгода-год сам арестованный аукнется или выбросят: «Без права переписки».*
(А. И. Солженицын. Архипелаг ГУЛаг. 1958–1973).

Исходная семантика данных глаголов, подразумевающая избавление от какого-либо ненужного объекта с помощью резкого движения руки отражается в реципиентной зоне в виде негативной коннотации, подчеркивающей резкость, быстроту и даже некоторую агрессивность действия. В целом, глаголы *выбрасывать* и *выбросить*, в семантике которых заложено представление о ненужности объекта, употребляются в семантической области глаголов речи значительно реже, чем глаголы *бросать* и *бросить*, не имеющие подобной семантики, но передающие только быстроту действия. Возможно, этот компонент значения в некоторой степени препятствует широкому функционированию глаголов каузации перемещения по воздуху в реципиентной зоне, что мы также сможем в дальнейшем наблюдать на примере каузативов другой группы.

К каузативам перемещения по воздуху, обладающим способностью к семантическому переходу в зону глаголов речи, примыкает глагол *выкинуть*, употребление которого связано с жесткими семантическими и синтаксическими ограничениями. Так данный глагол допускает единственное существительное *лозунг* для заполнения валентности содержания. Его видовой коррелят *выкидывать* не обладает даже этой способностью. Как представляется, само словосочетание *выкинуть лозунг* отсылает скорее не к собственно речевому акту произнесения лозунга, а к некоторой совокупности действий, приводящих к распространению данного призыва:

(5) *Епископ Русской Православной Церкви перед первой мировой выкинул лозунг: «Расступись! Русь идет!»* (Юлий Крелин. Почему у нас вечен конфликт истории и географии? // «Общая газета», 1996)

При этом глагол *выкинуть* не способен вводить ни прямую речь, ни придаточное предложение. Все это не дает нам достаточных оснований для того,

чтобы рассматривать глагол *выкинуть* как предикат, развивший способность к переходу в реципиентную зону глаголов речи. Вероятно, различия в поведении глаголов *выбросить* (*выбрасывать*) и *выкинуть* (*выкидывать*) связано с тем, что у глагола *выкинуть* развилось дополнительное значение с негативной коннотацией ‘проделать, устроить’, отсутствующее у глагола *выбросить*.

Крайне ограничены и употребления глаголов *отбрасывать* и *разбрасывать* в конструкции, характерной для глаголов речи. Они окказионально допускают существительное *слова* для заполнения валентности содержания. При этом употребление глагола *отбрасывать* подчеркивает резкость действия, а глагола *разбрасывать* — бессмысленность или неискренность речи:

(6) *И Брюсов, настойчивым методическим лаем, откусывая и отбрасывая слова: «Хотя я и не поклонник Ростана».*
(М. И. Цветаева. Герой труда (Записи о Валерии Брюсове) *1925)*

(7) *Поверьте мне, я не такой человек, чтобы разбрасывать слова, как пьяный матрос разбрасывает медяки».*
(Семен Липкин. Записки жильца (1962–1976)

Все эти данные не позволяют нам считать глаголы *отбрасывать* и *разбрасывать* предикатами с развитой способностью к семантическому переходу в исследуемую реципиентную зону.

Способность к переходу в семантическую зону глаголов речи отсутствует у предикатов *подбрасывать* (*подбросить*) и *подкидывать* (*подкинуть*), что отчасти может быть связано с многозначностью глагольной приставки *под-*, могущей одновременно означать движение вниз и движение вверх на небольшое расстояние [Плунгян 2001: 105], и, как следствие, развитием у данных глаголов в русском языке дополнительных значений ‘дать’ и ‘подложить’.

Таким образом, в русском языке для глаголов, производных от *бросить* (*бросать*) и *кинуть* (*кидать*), передающих семантику движения вниз и движения вверх на небольшое расстояние (*подбросить, подкинуть*), отторжения приближающегося объекта (*отбросить, откинуть*), распространения и рассеивания множественного объекта (*разбрасывать, раскидывать*) не характерен переход в семантическую область глаголов речи, что отчасти может быть связано с развитием у них других конкурирующих семантических переходов.

2.1.2. Группа *обронить*

Из данной группы глаголов, подразумевающих утрату контроля над объектом, наиболее частотным в рассматриваемой реципиентной зоне оказывается глагол *обронить*. Здесь им освоены все конструкции, доступные для каузативов перемещения по воздуху. Глагол *ронять* в значении ‘говорить, произносить’ нечасто встречается в современных текстах, тем не менее, как показывают данные НКРЯ, его можно встретить в текстах XIX века. Глагол *уронить* оказывается способным вводить прямую речь. Он также окказионально допускает дополнение *слово* (в корпусе найдено 2 подобных примера). Способность

вводить придаточное с союзом *что* у данного глагола отсутствует. Возможно, более высокая частотность глагола *обронить* по сравнению с *уронить* в рассматриваемых конструкциях связана с тем, что глагол *обронить* постепенно вытесняется глаголом *уронить* из исходной семантической зоны в реципиентную. Что касается глагола *проронить*, широко употребляющегося в качестве глагола речи, то он оказывается вытесненным из донорской зоны практически полностью. Наряду с функционированием в качестве своего рода глагола речи, в современном русском языке он допускает только дополнение *слеза*. В то же время, согласно словарю В. Даля, в XIX веке он являлся синонимом глагола *уронить* и имел значение ‘упустить наземь, потерять из числа многих вещей, а не одиночную’ [Даль 1994, т. 3: 505]. Все глаголы данной группы в базовом значении встраиваются в конструкцию $X_{\text{агенс}}$ V $Y_{\text{пациенс}}$. В исходном значении они подразумевают утрату агенсом контроля над пациенсом, в результате чего совершается падение, определенным этапом которого является перемещение по воздуху. В реципиентной зоне данные глаголы содержат в своей семантической структуре представление о ненамеренности действия и растерянности говорящего.

2.1.3. Глагол *отпустить*

До сих пор предметом нашего исследования являлись глаголы, подразумевающие пассивность объекта. Однако к уже рассмотренным предикатам примыкает глагол *отпустить*, допускающий наряду с активностью субъекта активность объекта. Данный глагол обладает способностью вводить прямую речь и придаточное изъяснительное с союзом *что*. Однако заполнять валентность содержания при данном глаголе в качестве прямого объекта может ограниченный набор существительных, в частности, *шутка*, *замечание* и т. д.:

(8) *Пока я дивился на него, не смея, конечно, улыбнуться или отпустить замечание, Дюрок подошел к стене между окон и потянул висячий шнурок.* (А. С. Грин. Золотая цепь. 1926)

Употребление данного глагола в качестве глагола речи влечет за собой следующие дополнительные смыслы: умышленный отказ от контроля над своей речью предполагает возможность сказать что-либо нелицеприятное, что подразумевает негативное либо презрительно-насмешливое отношение к собеседнику, предмету речи или ситуации в целом.

2.1.4. Глагол *выпалить*

В основе семантического перехода в структуре всех рассмотренных выше глаголов лежала метафора. В то же время при переходе глагола *выпалить* в семантическую зону глаголов речи задействованы механизмы метафоры и метонимии. Дело в том, что первоначально данный глагол имел значение ‘выжечь’ и мог принимать прямой объект (*выпалить поле*). Этот глагол содержал также представление о полном уничтожении объекта. Появление у данного глагола значения ‘выстрелить’ связано с тем, что первоначально для совершения этого

действия необходимо было поджечь порох в пушке. Представление о полном расходовании объекта при сохранении исходной аргументной структуры сохранилось в таких употреблениях, как *выпалить пушку* в значении ‘полностью израсходовать заряд’. В то же время у глагола *выпалить* появляются новые аргументы с семантической валентностью цели и/или источника, при этом представление о полном расходовании заряда отходит на второй план: *выпалить картечь в русский отряд*. Постепенно употребление данного глагола в новом значении задействует механизм метонимии, и происходит перестройка аргументной структуры глагола, которая принимает вид: $X_{агенс}$ V $Z_{цель}$ / $W_{источник}$. Первоначально единственный обязательный аргумент $Y_{пациенс}$, соотносясь неизбежно с зарядом орудия, перестает нуждаться в синтаксическом выражении и входит непосредственно в семантическую структуру глагола, становясь инкорпорированным актантом:

(9) *Гуров опять закричал: «Стой! » — и выпалил в воздух. Беглец и на этот раз проигнорировал его усилия и только увеличил скорость.* (Н. Леонов, А. Макеев. Ментовская крыша. 2004)

Метафорический переход данного глагола в семантическую зону глаголов речи основан на представлении о стремительности действия. В то же время при подобном употреблении синтаксическое выражение получает аргумент, утраченный в процессе метонимической перестройки семантической структуры глагола. В характерной для глаголов речи конструкции $Y_{пациенс}$ меняет свой таксономический класс и переходит в $Y_{содержание}$.

Итак, для глагола *выпалить* мы наблюдаем изменение аргументной структуры и перемещение из донорской зоны в реципиентную. Подобные изменения характерны для ребрендинга [Рахилина (ред.) 2010: 535]. Однако ребрендинг предполагает также изменение акционального класса глагола. В нашем случае мы сталкиваемся с изменением аргументной структуры глагола при сохранении акционального сходства донорской и реципиентной зон. В связи с этим можно утверждать, что в семантической структуре данного глагола мы наблюдаем два различных процесса: историческую метонимию и метафорический переход.

2.2. Каузативы перемещения, предполагающие выделение объекта через отверстие

До настоящего момента мы рассматривали глаголы, прототипическое значение которых предполагает действие, осуществляемое рукой. Перенос таких глаголов в семантическую зону глаголов речи широко распространен во многих языках: кит. *fā* ‘выстрелить’, тат. *ычкындыру* ‘отпускать, упускать, ляпнуть’ и т. д., хотя подобные употребления могут иметь иные коннотации, чем в русском, эта проблема остается за пределами нашего внимания в данной статье.

Сейчас мы хотели бы подробнее остановиться на вопросе о том, какие глаголы каузации перемещения, базовое значение которых не подразумевает

осуществления действия рукой, могут переходить в семантическую зону глаголов речи. В ряде языков, в частности, в адыгейском и абхазском, глаголы со значением ‘говорить’ (*Іон* и *аҳәара* соответственно) восходят к историческому корню *Іуы*, обозначавшему одновременно ‘рот’ и ‘край’ [Шагиров 1977:159–160], т. е. представление о речевом акте исторически как бы сближается с представлением о падении объекта. В русском языке подобная семантика представлена глаголами *выплевывать, изрыгать, извергать*. Эти предикаты содержат представление о выбрасывании некоторого преимущественно ненужного объекта изнутри через какое-либо отверстие, в частности, ротовое. Глагол *выплевывать* употребляется в функции глагола речи окказионально, а глаголы *изрыгать, извергать* накладывают определенные семантические ограничения на заполнение валентности содержания. Данную валентность способны заполнять такие существительные, имеющие явную негативную окраску, как *проклятия, оскорбления*. Таким образом, употребление данных предикатов в функции глаголов речи в русском языке ограничено очевидной негативной коннотацией.

2.3. Каузативы перемещения, предполагающие выделение некоторой субстанции всей поверхностью

Нужно отметить, что выделять некоторое вещество или субстанцию можно и всей поверхностью. Глагол именно с такой семантикой подвергается семантическому переходу в зону глаголов речи в иврите (*lifrot* ‘выбрасывать, извергать, испускать, излучать’[1]). В русском языке подобная стратегия не получает развития. Исследуемый нами в данной работе семантический сдвиг не наблюдается у глаголов *выделять, излучать*.

3. Заключение

Таким образом, относительно метафоры каузации перемещения по воздуху в семантической зоне глаголов речи в русском языке можно сделать следующие выводы. Подобный семантический переход характерен для глаголов бесконтактного перемещения, подразумевающих контроль только в начальной фазе действия (*бросать*). В русском языке особое место среди каузативов перемещения по воздуху, обнаруживающих способность к рассматриваемому семантическому переходу, занимают предикаты, в семантике которых заложено представление о некотором действии, осуществляемом рукой (*бросать*). Другая зона, включающая глаголы, содержащие представление о выталкивании объекта через некоторое отверстие, преимущественно ротовое (*изрыгать*),

[1] Данный глагол не подразумевает выталкивания объекта через какое-либо одно отверстие и, в отличие от русского глагола *извергать*, он не может использоваться для обозначения процесса извержения вулкана или акта рвоты.

в русском языке служит источником метафорического переноса значительно реже и оказывается ограниченной явными негативными коннотациями. У глаголов, подразумевающих выделение некоторой субстанции всей поверхностью (*выделять, излучать*), в русском языке не развивается способность к исследуемому метафорическому переходу, и представление о такой гипотетической донорской зоне мы получаем, рассматривая данные русского языка на фоне других языков. Способность каузативов перемещения по воздуху к переходу в семантическую зону глаголов речи может ограничиваться наличием в их семантической структуре других семантических переходов. В то же время переход в реципиентную зону глаголов речи непосредственно связан с исходным значением глагола. Необходимо отметить, что при осуществлении данного семантического перехода действуют ограничения на вид глагола, обусловленные сочетаемостью базовых глаголов речи *говорить* и *сказать*. В ходе рассматриваемого семантического перехода задействуются механизмы метафоры и метонимии.

Литература

1. *Даль, В.* Толковый словарь живого великорусского языка. М.: ТЕРРА, 1994.
2. *Зализняк, Анна А.* Русская семантика в типологической перспективе. М.: Языки славянской культуры, 2013.
3. *Кобозева, И. М.* О границах и внутренней стратификации семантического класса глаголов речи. / Вопросы языкознания. 1985. № 6, 95–103.
4. *Кустова, Г. И.* Типы производных значений и механизмы языкового расширения. М.: Языки славянской культуры, 2004.
5. *Никитина, Е. Н.* Видо-временные формы рамочных глаголов в русском нарративе XIX–XXI вв. / ACTA LINGUISTICA PETROPOLITANA. Труды Института лингвистических исследований РАН. Т. XI. Ч. 1. Категории имени и глагола в системе функциональной грамматики. СПб.: Изд-во «Наука», 2015. С. 825–847.
6. *Падучева, Е. В.* О семантическом инварианте видового значения глагола в русском языке. / Русский язык в научном освещении. 2004. № 2(8), 5–16.
7. *Плунгян, В. А.* Приставка *под-* в русском языке: к описанию семантической сети. / Московский лингвистический журнал. 2001, № 5(1), 95–124.
8. *Рахилина, Е. В.* (ред.) Лингвистика конструкций. М.: Азбуковник, 2010.
9. *Рахилина, Е. В., Лемменс, М.* Русистика и типология: лексическая семантика глаголов со значением ‘сидеть’ в русском и нидерландском. / Russian Linguistics. 2003. Vol. 27, № 3, 313–328.
10. *Шагиров, А. К.* Этимологический словарь адыгских (черкесских) языков. М.: Наука, 1977. Т. 2.

References

1. *Dal, V.* (1994), Explanatory Dictionary of the Live Great Russian language [Tolkovy slovar' zhivogo velikorusskogo yazyka], TERRA, Moscow.
2. *Kobozeva, I. M.* (1995), On the margins and inner stratification of the semantic field of speech act verbs [O granitsah i vnutrenney stratsifikatsii semanticheskogo klassa glagolov rechi], Problems of linguistics [Voprosy Yazykoznaniya], № 6, pp. 95–103.
3. *Kustova, G. I.* (2004), Types of derivative meanings and mechanisms of linguistic extension [Tipy proizvodnyh znacheniy i mekhanizmy yazykovogo rasshireniya], Yazyki slavyanskoy kultury, Moscow.
4. *Nikitina, E. N.* (2015), Tense and aspect forms of frame verbs in Russian narrative in 19–21 c. [Vido-vremennye formy ramochnyh glagolov v russkom narrative XIX–XXI vv.], ACTA LINGUISTICA PETROPOLITANA. V. XI. P. 1. Nauka, Saint-Petersburg, pp. 825–847.
5. *Paducheva, E. V.* (2004), On the semantic invariant of the aspectual verbal meaning in Russian [O semanticheskom invariante vidovogo znacheniya glagola v russkom yazyke], The Russian Language in a Scientific Light [Russkiy yazyk v nauchnom osveshchenii], № 2(8), pp. 5–16.
6. *Plungian, V. A.* (2001), The prefix *pod-* in Russian: on the description of the semantic network [Pristavka *pod-* v russkom yazyke: k opisaniyu semanticheskoy seti], Moscow Journal of Linguistics [Moskovskiy lingvisticheskiy zhurnal], № 5(1), pp. 95–124.
7. *Rakhilina, E. V.* (2010), Linguistics of constructions [Lingvistika konstruktsiy], Azbukovnik, Moscow.
8. *Rakhilina, E. V., Lemmens, M.* (2003), Russian studies and typology: lexical semantics of verbs meaning 'to sit' in Russian and Dutch [Rusistika i tipologiya: leksicheskaya semantika glagolov so znacheniem 'sidet'' v russkom i nidrlandskom, Russian Linguistics. Vol. 27, № 3, pp. 313–328.
9. *Shagirov, A. K.* (1977), Etymological Dictionary of Adyghe (Circassian) languages [Etimilogichesky slovar' adygskih (cherkesskih) yazykov. V. 2. Nauka, Moscow.
10. *Zaliznyak, Anna A.* (2013), Russian semantics from the typological perspective [Russkaya semantika v tipologicheskoy perspektive], Moscow.

К ПРОБЛЕМЕ СОПОСТАВИТЕЛЬНОГО АНАЛИЗА ПРОСОДИИ: ОДЕССКИЙ РЕГИОНАЛЬНЫЙ ВАРИАНТ РУССКОГО ЯЗЫКА VS. РУССКАЯ РАЗГОВОРНАЯ НОРМА[1]

Янко Т. Е. (tanya_yanko@list.ru)

Институт языкознания РАН, Москва, Россия

«Одесскому акценту» посвящена большая литература, в которой исследуются сегментные фонетические, синтаксические, лексические и стилистические особенности языка Одессы. Между тем особенности одесской просодии практически не исследованы. Встает задача сопоставления просодии русского языка Одессы с просодией русской разговорной нормы. Просодическая норма понимается здесь как следование системе интонационных конструкций, описанных Е. А. Брызгуновой, плюс соответствие этих конструкций функциям, которые они выполняют в нормативном дискурсе. Таким образом, интонационные конструкции анализируются не только с точки зрения различительных признаков, но и в связи с дискурсивными значениями, которые они манифестируют. Просодия, таким образом, исследуется на различных языковых уровнях: на уровне различительных признаков акцентов (моделей изменения частоты тона и интенсивности, наложения изменений частоты на сегментный материал), на уровне акцентов — интонационных конструкций в духе Е. А. Брызгуновой — и на уровне функций просодии в дискурсе. Разработан исследовательский массив звучащей «одесской» речи: это рассказы одесситов о городе, анекдоты, записи кулинарных телепередач, интервью с одесситами. При обработке звучащих материалов использована компьютерная программа Praat.

Ключевые слова: интонация, коммуникативная структура, дискурсивные значения, одесский акцент, русская разговорная норма, русский язык, корпусный метод, звучащая речь, Одесса

[1] Исследование выполнено при поддержке Российского научного фонда (РНФ), проект № 14-28-00130.

CONTRASTIVE ANALYSIS OF PROSODY: ODESSA REGIONAL RUSSIAN VS. STANDARD RUSSIAN

Yanko T. E. (tanya_yanko@list.ru)

Institute of lingustics, Moscow, Russia

There is a considerable body of literature referring to segmental phonetic, syntactical, lexical, and stylistic parameters of Odessa accent. However, the prosodic peculiarities of spoken Odessa Russian are seriously underestimated, particularly if we take into consideration the role Odessa language plays in the Russian culture. This paper is aimed at contrasting the Odessa prosody to the prosody of the standard Russian spoken language. The Russian standard prosody, as it is viewed here, is the system corresponding to the inventory of intonational constructions recognized by E. A. Bryzgunova, including their functions in the standard Russian spoken discourse. Prosody is thus referred to not only as a system of distinctive features of the spoken language but also as the basic means of manifesting the communicative meanings: the illocutionary force, the contrast, the discourse continuity. Prosody is analyzed at the level of distinctive features of pitch accents (such as the pitch movement patterns, patterns of the pitch alignment with the text, intensity), at the level of integral pitch accents as they are represented by E. A. Bryzgunova, and at the level of the pitch accents as manifestations of the communicative meanings. For investigation, a minor working corpus of Odessa speech recordings was set up. The corpus consists of interviews with the speakers of Odessa Russian, short stories about Odessa told by the citizens, cooking recipes, jokes, and funny stories. The software programs Praat and Speech Analyzer were used in the process of analyzing the sounding data. The results presented here are exemplified by frequency and intensity tracings of records from Odessa speech oral corpus.

Key-words: prosody, pitch accents, Odessa accent, standard spoken Russian, corpus methods, communicative structure, spoken corpora, Odessa

Я пишу с акцентом,
читаю с акцентом
и меня с акцентом слушают
(М. Жванецкий).

«Одесскому акценту» посвящена большая литература, в которой исследуются сегментные фонетические, синтаксические, лексические и стилистические особенности языка Одессы, см. [Вершик 2003; Степанов 2004; Мечковская 2006; Kabanen 2008] и цитированные там публикации. Между тем особенности одесской просодии практически не исследованы. В работах, посвященных языку Одессу, говорится об особой восходяще-нисходящей интонации (rise-fall intonation), которая присуща языку идиш и которая, по мнению авторов, характерна также для русского языка

Одессы, ср., например, [Вершик 2003: 143][2]. Между тем, во-первых, понятие восходяще-нисходящего контура представляется недостаточно конкретным в применении к описанию просодии языка в целом и, во-вторых, влияние идиша на русский язык Одессы, которое мы, в принципе, не подвергаем сомнению, с теоретической точки зрения должно служить предметом специального научного доказательства, равно как и другие влияния, которым, несомненно, подвергался язык Одессы. Такая задача в настоящей работе не ставится. Цель данной работы — сопоставление одесской просодии с просодией русской разговорной нормы (иначе — наддиалектного русского), что должно пролить свет на особенности одесского акцента.

В ходе анализа мы опирались на данные эксперимента, который состоял в следующем. Звучащий отрезок из речи носителя одесского регионального варианта русского языка, который в соответствии со слуховым экспертным заключением содержал в себе черты одесского просодического колорита, дублировался носителем литературного наддиалектного варианта русского языка в лабораторных условиях. Нормативная артикуляция по возможности максимально отражала коммуникативное членение, которое определяло просодию оригинала. При этом мы исходили из гипотезы о том, что перед нами различные, но сопоставимые просодические системы, как по различительным просодическим признакам, так и по функциям. В частности, мы вообще не останавливались на сопоставлении значений такого важного просодического параметра, как выбор словоформ-носителей акцентных пиков предложения: мы исходили из того, что это не просто сопоставимый выбор, а такой, при котором носители акцентов в обоих диалектах совпадают. Так, в предложении *Есть у нас такой двадцать восьмой трамвай…* носитель акцентного пика — словоформа *трамвай* в обоих диалектах, при том что сами акценты, как будет показано ниже на примере (1), могут иметь свои отличия. Оба варианта — «одесский» и нормативный — сравнивались, и системные просодические отличия и совпадения фиксировались описательно или в специально подобранных терминах. Так, для описания одесской просодии использовалась нотация, основанная на нотации Е. А. Брызгуновой [1982, 96–118], и автосегментной просодической транскрипции Дж. Пьерхамберт [Pierrehumbert 1980], разработанной для описания американского диалекта английского языка. В частности, использовался вариант автосегментной нотации, адаптированный к просодии немецкого языка [Grice, Baumann, Benzmüller 2005]. Использование английской и немецкой нотации объясняется фонетической близостью некоторых «одесских» и западноевропейских просодий.

Одесский акцент не всегда в равной степени присутствует в речи одесситов, которые владеют наддиалектной нормой и используют ее в своей речи. Есть более устойчивые инвариантные черты, которые проявляются практически всегда, и есть более подвижные, которые педалируются при рассказывании анекдотов и непринужденном цитировании чужой речи, но преодолеваются в более формальной обстановке. Таким образом, стоявшая перед нами задача осложнялась принципиальной неоднородностью имеющихся данных.

---

[2] На восходяще-нисходящий контур идиша указывают и другие авторы, см. список цитированной литературы по этому вопросу в статье [Светозарова 2007: 233].

В разделе 1 рассматриваются фонетические особенности «одесских» просодий (мелодические модели, способы наложения изменений частоты основного тона на сегментный материал, модели просодической и сегментной редукции), в разделе 2 —интонационные конструкции одесского диалекта в фонологическом аспекте и их наддиалектные корреляты, раздел 3 посвящен функциям «одесских» просодий в звучащей речи.

1. Интонационная фонетика. Различительные признаки акцентов и высказываний

1.1. Поздний тайминг

Рассмотрим параметры одесской просодии на примерах. Обратимся к предложению (1).

(1) *Есть у нас такой 28* ***трамвай****, который едет по улице* ***Пантелеймоновской*** *напротив* ***Привоза***[3].

Рис. 1 отражает график изменения частоты основного тона в герцах (или тонограмму, жирная прерывистая кривая) и график изменения интенсивности звучания в децибелах (тонкая непрерывная кривая, которая накладывается на тонограмму). Верхняя панель фиксирует параметры оригинала (одесскую речь), на нижней панели — представлены значения параметров нормативного коррелята. Графики получены с помощью компьютерной системы анализа устной речи Praat: кроме того, что графики значений основных параметров — это одно из базовых — наряду со слуховым — средств анализа, это и наиболее удобное средство передачи на бумаге параметров звучащей речи.

Одесский оригинал демонстрирует два значимых подъема частоты на словоформах *трамвай* и *Пантелеймоновской* и финальное падение на *Привоза*. Мы сейчас не останавливаемся на проблеме трактовки подъемов как показателей незавершенности повествования или как темы сообщения, противопоставленной реме. (Рема в данном случае — это финальное обстоятельство *напротив Привоза* с акцентоносителем словоформой *Привоза*.) При использованном в примере (1) типе повествовательной стратегии тема и дискурсивная незавершенность в рамках сложного предложения формально не различаются — ни по типу акцента, ни по способу выбора акцентоносителя — при том, что в русской речи существуют стратегии, позволяющие сохранить это различие, см. об этом [Янко 2008: 128–155]. В данном случае концептуальное различие между темой предложения и дискурсивной незавершенностью непринципиально.

Сравнение акцентов на словоформе *трамвай* позволяет сделать следующие выводы. В «одесском оригинале» предударный слог существенно сильнее

[3] Привоз — продуктовый рынок в Одессе.

редуцирован, чем при наддиалектном прочтении: интенсивность и длительность предударного слога по отношению к ударному слогу здесь ниже, чем в наддиалектной форме. Кроме того, имеется отличие и в сегментной реализации. Наддиалектная норма реализуется с а-образным звуком [ʌ], а одесский вариант — с редуцированным звуком «шва» [ə], что опять же иллюстрирует бóльшую степень предударной редукции в языке Одессы и, как следствие, бóльшую выделенность ударного слога. Заударные слоги здесь отсутствуют. Отметим также более узкий диапазон подъема, который фиксируется при «одесской» артикуляции словоформы *трамвай*.

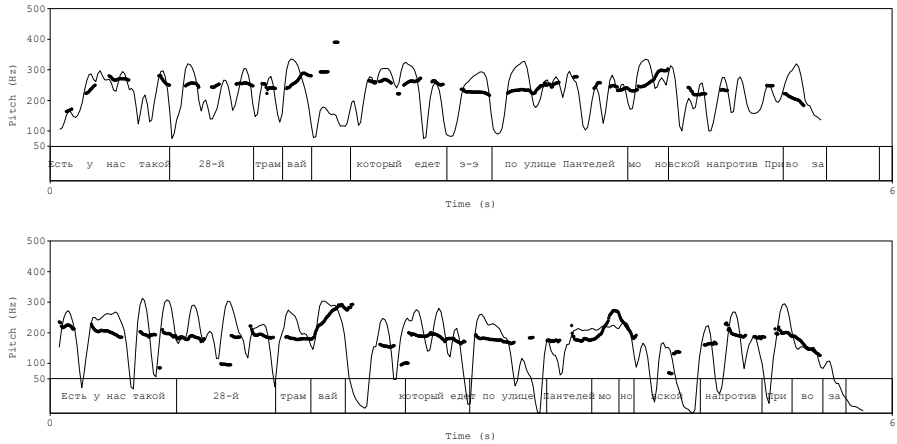


Рис. 1. Параметры исходной (одесской, верхняя панель) и стандартной (нижняя панель) версий предложения (1)

Аналогично на словоформе *Пантелеймоновской* предударный слог в одесском оригинале имеет более существенную степень редукции, как по отношению к ударному слогу, так и по отношению к наддиалектной норме, кроме того, он отмечен суженным мелодическим диапазоном. Наблюдаются и существенные различия на заударных слогах. А именно: в наддиалектной норме представлено рельефное заударное понижение частоты на первом заударном слоге. Перед нами типичная реализация ИК-3 [Брызгунова 1982: 111–114]. Между тем в одесском варианте подъем на *Пантелеймоновской* продолжается и на первом заударном слоге. Просодическая конструкция «подъем на ударном слоге плюс продолжающийся подъем на заударном» в литературном русском представлена редкими спорадическими употреблениями и в списке интонационных конструкций Е. А. Брызгуновой не значится. Для немецкого же языка эта конструкция, напротив, весьма характерна, и для нее предусмотрен специальный транскрипционный знак L*H-^H% [Grice, Baumann, Benzmüller 2005], где фрагмент L*H (движение частоты от низкой точки (L), к высокой (H)) обозначает подъем на ударном слоге, а фрагмент -^H% — градуальный подъем на заударном. Знак «^» указывает на то, что подъем на заударном начинается от того уровня частоты, который достигнут на ударном слоге, ср. (2):

(2) Нем. *Ich war in **Frankreich**. Dann war ich einen Monat in **China**…*
‘Я была во Франции. Потом я была один месяц в Китае…’

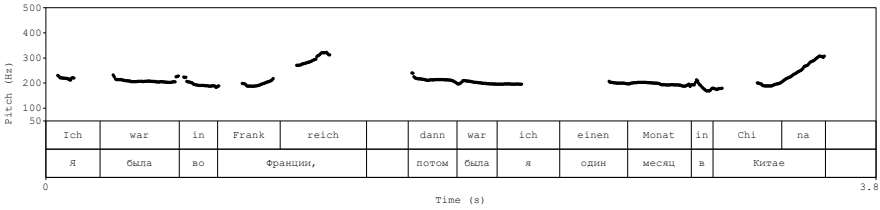


Рис. 2. Тонограмма примера (2)

Тонограмма примера (2) демонстрирует два градуальных подъема на *Frankreich* и *China*. В немецком языке такой акцент обозначает особый вид незавершенности дискурса, а именно: т. н. «рассказ по порядку». Рассказ по порядку предполагает, что повествование выстроено рассказчиком в соответствии с порядком, в котором происходили события, — в примере (2) — это посещение стран, где побывал говорящий во время летних каникул. Пример заимствован из работы [Палько 2008]. Близость частотных показателей словоформы *Пантелеймоновской* с учетом способа наложения на ударный и заударный слоги в языке Одессы и немецких словоформ в примере (2) налицо. Открытым остается вопрос о фонологическом статусе этой конструкции в одесском диалекте. В немецком языке в соответствии с нашими наблюдениями акцент L*H-^H%, как уже говорилось выше, служит средством выражения одного из типов незавершенности — рассказа по порядку. В одесском же языке у этой конструкции подобной инвариантной функции (или функций) мы не наблюдаем. Возможно, при дальнейших исследованиях такие функции и будут обнаружены. Однако на настоящем уровне исследований мы не исключаем, что данная конструкция может служить результатом одесского «сдвига» фокуса артикуляции вправо, или т. н. позднего тайминга. Поясним, что при позднем тайминге предельные значения просодических параметров смещаются на вторую часть ударного слога словоформы-акцентоносителя или на заударный слог. В таком случае в Одессе конструкция L*H-^H% может не иметь фонологического статуса, оставаясь фонетическим вариантом одного из восходящих акцентов, например, акцента типа ИК-3, т. к. при позднем тайминге восходящая артикуляция может захватывать заударный слог.

Пример (1) служит наглядной иллюстрацией феномена позднего тайминга, который характеризует способ наложения частотных показателей на сегментный материал. К симптоматике позднего тайминга в одесском диалекте мы относим «одесскую» редукцию предударного слога (сегментная редукция, уменьшение длительности звучания, низкая интенсивность), суженный мелодический диапазон ударной части, сдвиг вектора основного движения тона на первый заударный слог, отсутствие характерного для наддиалектной нормы «затухания» амплитуды колебаний на заударных слогах акцентоносителей, а также продленное, интенсивное и рельефное по форме вектора изменения частоты произнесения заударных, что говорит о не сравнимой с наддиалектной нормой энергетике

артикуляции заударных слогов. Отсутствует на заударных слогах и сегментная редукция. Нисходящий акцент на словоформе *Привоза*, тоже демонстрирует сдвиг интенсивности, амплитуды колебаний и скорости изменения частоты «вправо» в одесской артикуляции по сравнению с наддиалектной. Однако в силу сдвига «вправо» реализация нисходящего тона по сравнению с восходящим, наоборот, оказывается более рельефной по причине большего «энергетического» наполнения второй части ударного и заударного — тоже нисходящего — тона.

Аналогично, тонограмма примера (3) из одесского диалекта на рис. 3 демонстрирует крутое падение частоты на ударном слоге акцентоносителя ремы словоформы *моря* и беспрецедентно длительный (для уха носителя наддиалектного варианта русского языка) заударный слог *-ря*, сегментная редукция на котором также отсутствует. В наддиалектной норме финальное падение существенно более «плоское», чем в одесском диалекте, оно характеризуется затуханием интенсивности, амплитуды колебаний и, соответственно, потерей звучности.

(3) *Вы находитесь в северо-западной части Черного* ***моря***.

Рис. 3. Тонограмма примера (3)

1.2. Длительность заударных слогов

Минимальная заударная редукция и эффект позднего тайминга создают возможность повышения длительности заударных слогов, которая служит характерной чертой одесского диалекта. О минимальной заударной редукции и позднем тайминге свидетельствует длительность заударного слога *-ва* словоформы-акценоносителя *Петрова* в примере (4); она составляет 800 мсек (ударного — ок. 300 мсек).

(4) *Шо вы не читали Ильфа и* ***Петрова-а-а****?*

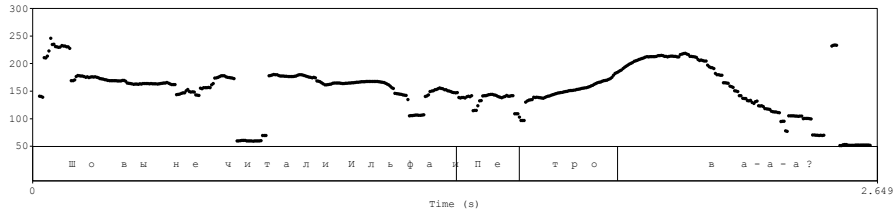


Рис. 4. Тонограмма примера (4)

1.3. Downstepping

В одесском диалекте наблюдается отсутствующий в наддиалектной норме феномен последовательного повтора в рамках предложения одной и той же просодической конструкции. В англоязычной традиции такой повтор называется downstepping [Pierrehumbert 1980: 139–176]. Он характеризуется тем, что каждая n+1-пая встречаемость повторяющейся конструкции произносится на более низком среднем уровне, чем n-ая встречаемость. Downstepping формирует просодический контур предложения в целом или его основного коммуникативного компонента — ремы. Иначе говоря, downstepping — это градуальное повторение конструкции на каждой последующей тактовой группе на уровень ниже предыдущего этапа вплоть до нижнего значения тона голоса данного говорящего. Downstepping иллюстрируется примерами (5)–(6):

(5) *Я думал, шо вы хуже.*

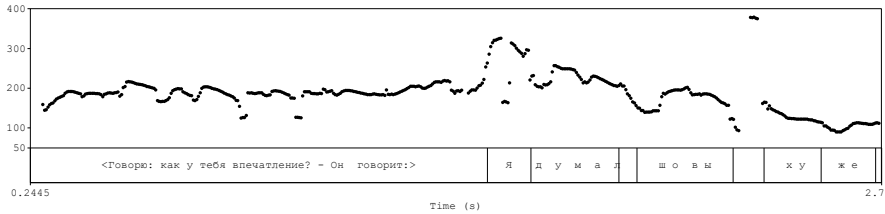


Рис. 5. Тонограмма примера (5)

Пример (5) иллюстрирует downstepping в одесском диалекте. На первом ударном *я* наблюдается падение частоты основного тона, на ударном слоге словоформы *думал*, фиксируется аналогичное падение, которое продолжается и на заударном слоге, фрагмент *шо вы* произносится практически безударно, но форма кривой сохраняется, на финальной словоформе-акцентоносителе *хуже* говорящий достигает нижнего предела своего частотного диапазона. С перцептивной точки зрения downstepping имеет монотонное звучание нисходяще-скандирующего характера, ср. англ. пример:

(6) Англ. *I stayed in the US Navy seventeen years and ten months.*

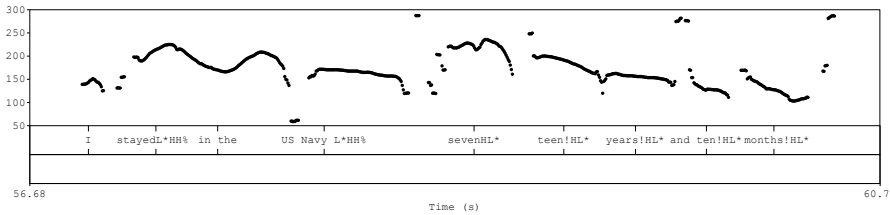


Рис. 6. Тонограмма примера (6)

Пример (6) демонстрирует повтор нисходящеего (HL*, по [Pierrehumbert 1980]) движения тона на ударных слогах словоформ *seventeen* (сложное слово артикулируется здесь с двумя акцентами: с акцентом на *seven* и с акцентом на *teen*), *years*, *ten* и *months*. Для обозначения эффекта downstepping'а здесь используется восклицательный знак (по [Pierrehumbert 1980], ср. также [Odé 2002]).

1.4. Upstepping

Явление upstepping'а аналогично downstepping'у, но связано, наоборот, с градуальным подъемом каждого фонетического слова на уровень, выше предыдущего, вплоть до высшей точки голоса данного говорящего. По нашим наблюдениям upstepping представлен и в одесском диалекте:

(7) *Что вы хотите еще рассказать вам в Одессе?*

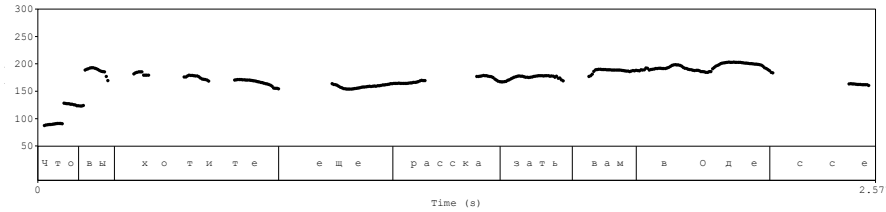


Рис. 7. Тонограмма примера (7)

В примере (7) наблюдается ряд подъемов тона на фонетических словах *еще*, *рассказать вам* и *в Одессе*.

Для сравнения, градуальное развертывание ритмических групп вверх и вниз по шкале изменения частоты основного тона характерно для коротких предложений датского языка, описанных в работах Н. Грённум [Groennum 1998: 298–310]. Downstepping и upstepping — это пример т. н. интегральных просодий, характеризующих предложение в целом, они представляют собой один из видов деклинации и инклинации; об интегральных просодиях см. [Кодзасов, Кривнова 2001: 384, 396].

1.5. «Высокий конец» и обратная адаптация

Одним из средств выражения значения несогласия и контраста в одесском диалекте служит высокое расположение акцентоносителя ремы, занимающего одну из финальных позиций в предложении. Соответственно, акцентоноситель темы принимает в таком предложении противоположный вектору ремы и несвойственный теме нисходящий акцент, который обусловлен высоким расположением акцентоносителя ремы на тональной шкале.

(8) *Там, где не понимают **русского**, я **рисую**.*

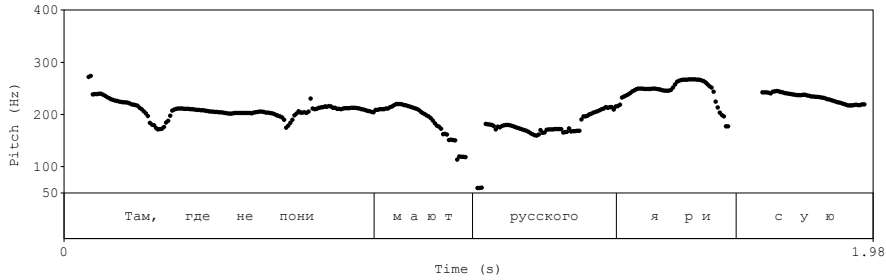


Рис. 8. Тонограмма примера (8)

На акцентоносителе ремы словоформе *рисую* наблюдается высокий — не характерный для простой ремы, которая не осложнена значением контраста или несогласия, — тон. Соответственно, частоты акцентоносителя темы — словоформы *русского* — низкие, слабо нисходящие. Эффект обратной адаптации не служит отличительной особенностью именно одесского диалекта. Характерное «одесское» явление, которое иллюстрируется примером (8), состоит в падении тона, маркирующего тему, а падение, в свою очередь, обусловлено двумя факторами: специфическим одесским «высоким концом» предложения и обратной адаптацией.

Фонетические просодические особенности одесского диалекта состоят в следующем: поздний тайминг, суженный мелодический диапазон при реализации восходящих тональных акцентов, и, наоборот, рельефная артикуляция нисходящих акцентов, сильная степень предударной редукции, отсутствие заударной редукции, заударные растяжки, down- и upstepping.

2. Интонационная фонология. Некоторые различия одесской и наддиалектной систем интонационных конструкций в терминах Е. А. Брызгуновой

2.1. Одесский аналог ИК-1

При маркировании ремы установленная выше тенденция одесской просодии к смещению интонационной кульминации «вправо» приводит к тому, что аналог ИК-1 артикулируется в среднем в бóльших, чем в наддиалектном русском, диапазоне частот. Соответственно, конструкция ИК-1 (по [Брызгунова 1982: 109]), которая представляет своего рода «затухающее» движение тона — и по скорости изменения частоты, и по амплитуде колебания, и по интенсивности — в одесском диалекте имеет более рельефный аналог, более близкий к нормативной конструкции ИК-2 [Брызгунова 1982: 109–111]. В примере (9) наблюдаются два значимых падения частоты: на словоформах *армии* и *Преображения*. Обоим падениям предшествует небольшой «заход» частоты вверх внутри ударного слога, характерный

и для наддиалектного ИК-2. Этот подъем служит для «набора высоты» и ведет к явлению позднего тайминга и в наддиалектном русском. Заударные слоги демонстрируют рельефное падение частоты, продленное звучание и отсутствие сегментной редукции, характерные уже для одесского диалекта.

(9) … *на бывшей улице Советской* ***армии****, теперь —* ***Преображения****.*

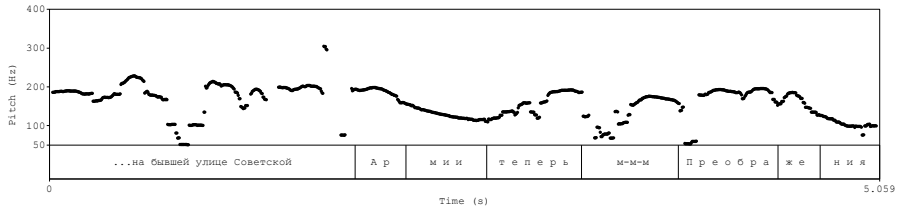


Рис. 9. Тонограмма примера (9)

2.2. Одесский аналог ИК-3

Одесский коррелят ИК-3 отличается сдвигом пика частоты основного тона «вправо», тенденцией к сужению мелодического диапазона, сегментной редукцией предударных слогов при их наличии (ср. примеры (1) и (4)) и отсутствием заударной редукции. Не входя в детали, проиллюстрируем эти параметры примером (10). На верхней панели рисунка (10) представлена тонограмма одесского варианта вопроса (10), на нижней — наддиалектного (см. [Брызгунова 1982: 111–114]).

(10) *Поняли?*

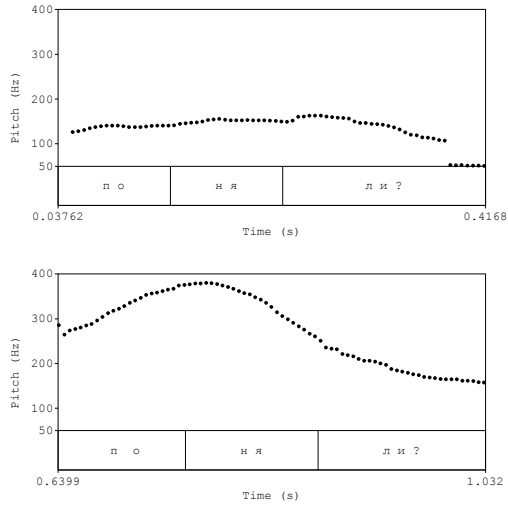


Рис. 10. Сопоставительная тонограмма двух реализаций примера (10)

Рис. 10 иллюстрирует сдвиг максимума частоты «вправо» в одесском диалекте по сравнению с наддиалектной нормой.

2.3. Одесский аналог ИК-6

Аналогично ИК-3, одесский коррелят ИК-6 демонстрирует редукцию, включая сегментную, на предударном слоге, подъем с небольшим перепадом частот на ударном слоге, сопровождающийся растяжкой, а также отсутствие редукции и растяжку на заударных слогах, если они имеются. Сравним две реализации словоформы *колба́сы* из примера (11). Словоформа *колба́сы* удачно сочетает в себе присутствие предударного и заударного слогов.

(11) *…в павильон, где продаются* ***колба-а-сы-ы****, как говорится,* ***ветчинка*** *такая вот…*

На верхней панели рис. 11 можно наблюдать все особенности «одесской» артикуляции ИК-6, в частности, сверхрастяжку на заударном слоге словоформы *колба́сы*: здесь ударный слог имеет длительность 400 мсек, заударный — 582 мсек. На нижней панели представлена нормативная реализация ИК-6 с акцентоносителем *колбасы*. Общее время артикуляции здесь на 50 мсек меньше, чем в одесском варианте, изменения частот имеют более рельефный характер, предударная редукция выражена существенно слабее, чем в одесском варианте, заударная — сильнее.

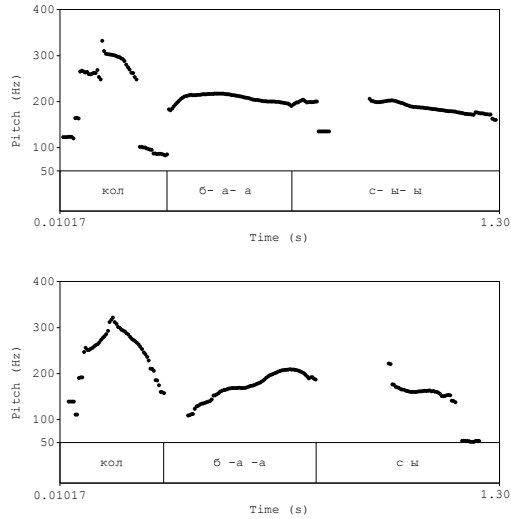


Рис. 11. Сопоставительная тонограмма артикуляции словоформы *колбасы*

Итак, основные различия между интонационными конструкциями в духе Е. А. Брызгуновой сводятся к различиям между ИК-1, ИК-3 и ИК-6 и их «одесскими» аналогами. Что касается ИК-2, то эта русская просодическая конструкция с фонетической точки зрения в существенной мере занимает нишу ИК-1 в силу сдвига кульминации артикуляции в одесском диалекте на вторую часть ударного слога и заударный слог. Аналоги для ИК-5 и ИК-7 как периферийные на данном этапе анализа остались без рассмотрения. Что касается ИК-4, то существенных произносительных отличий между стандартной и одесскими реализациями мы не наблюдаем. Имеются некоторые функциональные отличия, которые рассматриваются в следующем разделе. Фонологический статус акцента с подъемом на ударном слоге плюс градуальный подъем на заударных L*H-^H% (примеры (1)–(2)) остается открытым.

3. Функции интонационных конструкций (акцентов) в высказывании и в дискурсе

Жанр небольшой статьи позволит нам рассмотреть не все, а только некоторые функции акцентов.

Основная функция акцента ИК-1 (в его одесской модификации) — маркирование ремы. В качестве примера можно привести просодическое оформление акцентоносителя ремы словоформы *Привоза* в примере (1). При высоком расположении на шкале частот акцентоносителя ремы в композиции с контрастом, акцент типа ИК-1 может маркировать и тему, ср. пример (8) с акцентоносителем темы *русского*.

Акцент типа ИК-3 (с учетом «одесских» сегментных и просодических особенностей, которые в этом разделе уже не комментируются) манифестирует *да-нет*-вопрос (ср. пример (10) выше), тему (пример (12)) и незавершенность дискурса (13).

В (12) акцент типа ИК-3 фиксируется на акцентоносителе словоформе *назад* темы *приблизительно десять дней тому назад*.

(12) *Приблизительно десять дней тому* ***назад*** *я приглашал на спектакль ребят из Нацгвардии*.

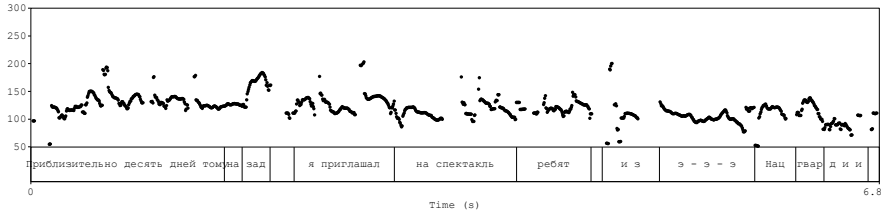


Рис. 12. Тонограмма примера (12)

Пример (13) демонстрирует использование акцента типа ИК-3 в режиме незавершенного повествования. Акцентоноситель значения незавершенности здесь словоформа *каникулы*. Функциональных отличий от стандартного употребления мы не наблюдаем.

(13) *Как раз зимние* ***каникулы*** *у студентов были…*

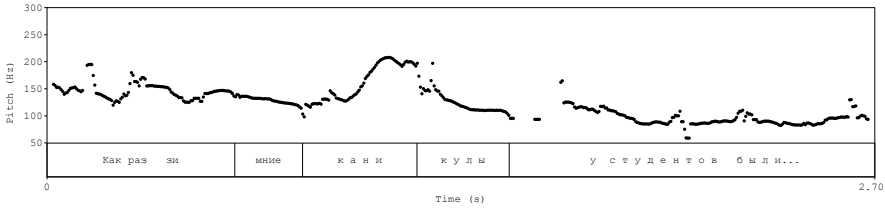


Рис. 13. Тонограмма примера (13)

Если вернуться к просодической конструкции L*H-^H%, которая обсуждалась на примере предложения (1), то предварительно можно предположить, что она представляет собой фонетический вариант акцента ИК-3 в условиях ярко выраженного позднего тайминга, который проявляет себя в захвате подъемом, образующим этот акцент, не только ударного, но и заударных слогов. Решение о нефонологическом статусе этой конструкции возникает в связи с тем, что отдельных функций, отличных от стандартной реализации ИК-3 с падением на заударных, у L*H-^H% мы не обнаруживаем. Мы не исключаем, что в дальнейшем появятся новые данные, которые изменят это решение. Обратимся к примеру.

(14) *Естественно, не всегда распространялась полностью* ***продукция****…*

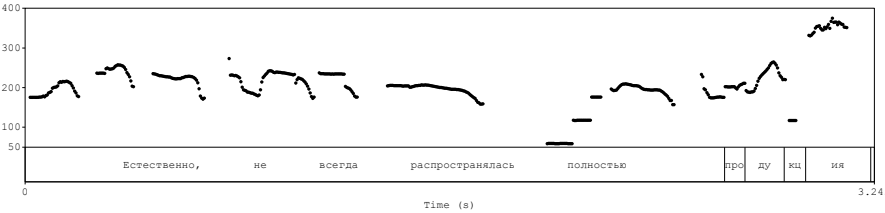


Рис. 14. Тонограмма примера (14)

На рис. 14 мы наблюдаем высокие заударные, которые отличают реализацию словоформы *продукция*. Это ощущается как ярко выраженный «акцент», легко уловимый ухом говорящего на наддиалектном русском.

У акцента типа ИК-4 в целом те же функции, что и в наддиалектном русском [Янко 2008: 209–218]: «рассказ по порядку», ответ и вопрос с «вызовом», приветствие, прощание (в режиме скорой встречи). Эти функции связаны с семантикой

сопоставления, противопоставления, встраивания в упорядоченную последовательность объектов или событий. Одесская специфика ИК-4 определяется бóльшей частотностью акцента типа ИК-4 по сравнению с наддиалектной речью. Эту черту можно объяснить как бóльшим полемическим задором одесситов, так и западноевропейским влиянием, где акцент типа ИК-4 служит немаркированным показателем дискурсивной незавершенности и вопроса, ср. эффект т. н. «рассказа по порядку» в (15):

(15) *Одна дама пошла покупать на* ***Привоз****…*

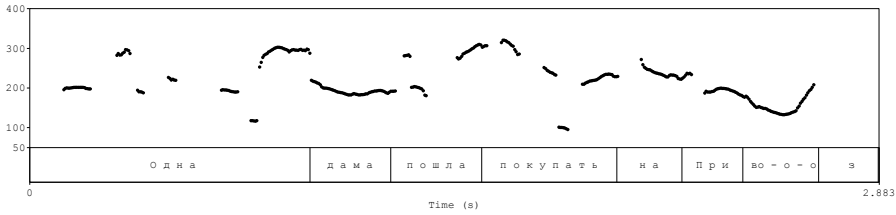


Рис. 15. Тонограмма примера (15)

На ударном слоге словоформы *Привоз* фиксируется рельефное нисходяще-восходящее движение тона, которое в данном случае говорит о незавершенной цепи событий, которые произойдут с дамой на Привозе.

Анализ функций одесской просодии в данном разделе был намечен в направлении от акцентов — к функциям. Задача же полного и системного описания средств выражения коммуникативных значений и функций — иллокутивных значений, значений, модифицирующих компоненты речевых актов, таких как контраст и эмфаза, а также значений, поддерживающих связность звучащего дискурса, — требует большего объёма, чем допускает жанр небольшой статьи. Здесь эта задача решена не была, это вопрос будущего.

***

В соответствии с нашей гипотезой загадка одесской просодии состоит в сдвиге кульминации коммуникативно релевантного акцента «вправо» от ударного слога словоформы-акцентоносителя. С этим параметром связана более существенная, чем в наддиалектном русском, степень редукции первого предударного слога, который отличается слабой интенсивностью, сокращенным временем звучания и сегментной редукцией, а также сдвиг максимума интенсивности и частотного пика на вторую часть ударного или на заударный слог. Заударный слог характеризуется слабой редукцией или ее отсутствием и допускает беспрецедентные для уха носителя наддиалектного русского повышения длительности звучания. В одесском диалекте имеются особые виды деклинации и инклинации интегральных движений тона, при которых наблюдаются последовательные повторы интонационных

конструкций, каждая из которых имеет среднюю частоту ниже (при деклинации) или выше (при инклинации), чем предыдущие. Удается установить фонетические и функциональные соответствия между конструкциями наддиалектного русского и их одесскими коррелятами. Более детальный анализ одесской просодии составляет перспективу настоящего исследования.

Литература

1. *Брызгунова Е. А.* (1982) Интонация, Русская грамматика, том 1, Наука, Москва, сс. 103–118.
2. *Вершик А.* (2003) О русском языке евреев, Die Welt der Slaven, Vol. XLVIII, pp. 135–148.
3. *Кодзасов С. В., Кривнова О. Ф.* (2001) Общая фонетика. РГГУ, М.
4. *Мечковская, Н. Б.* (2006) Русский язык в Одессе: Вчера, сегодня, завтра, Russian Linguistics, Vol. 30, No. 2. pp. 263–281.
5. *Палько М. Л.* (2008) Интонация незавершенности текста в немецком языке в сопоставлении с русским, Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог-2008». — М.: РГГУ, 2008. — Вып. 7 (14). — С. 416–419, available at: http://www.dialog-21.ru/digests/dialog2008/materials/html/65.htm
6. *Светозарова Н. Д.* (2007) Краткий очерк фонетики языка идиш в сравнении с русским и немецким, Acta Linguistica Petropolitana, Том 3, Часть 1, сс. 224–239.
7. *Степанов Є. М.* (2004) Російське мовлення Одеси, Астропринт, Одеса.
8. *Янко Т. Е.* (2008) Интонационные стратегии русской речи в сопоставительном аспекте, Языки славянских культур, Москва.
9. *Grice M., Baumann S., Benzmüller R.* (2005). German Intonation in Autosegmental-Metrical Phonology, Prosodic Typology: The Phonology of Intonation and Phrasing. Oxford University Press, available at: http://www.coli.uni-saarland.de/publikationen/softcopies/Grice:19xx:GIA.pdf
10. *Groennum N.* (1998) Fonetik og Fonologi. Almen og Dansk. Akademisk Forlag, Koebenhavn.
11. *Kabanen I.* (2008) Введение в особенности одесского языка. Helsinki, на: http://www.helsinki.fi/venaja/opiskelu/graduja/kabanen.pdf
12. *Odé C.* (2002) Перспективы описания и транскрипции русской интонации в корпусах звучащих текстов, Проблемы и методы экспериментально-фонетических иссследований. К 70-летию профессора Л. В. Бондарко, Спб., сс. 209–214.
13. *Pierrehumbert J.* (1980) The Phonology and Phonetics of English Intonation, MIT PhD Dissertation, available at: http://faculty.wcas.northwestern.edu/~jbp/publications/Pierrehumbert_PhD.pdf

References

1. *Bryzgunova E. A.* (1982) Intonation [Intonatsiya], Russian Grammar [Russkaya grammatika]. Vol. 1, Nauka, Moscow, pp. 98–118.
2. *Vershik A.* (2003) On the Russian of the Jews [O russkom yazyke evreev], Die Welt der Slaven, Vol. XLVIII, pp. 135–148.
3. *Kodzasov S. V., Krivnova O. F.* (2001) General phonetics [Obshchaya fonetika]. Russian State University for Humanities, Moscow.
4. *Kabanen I.* (2008) Introduction to the parameters of the language of Odessa [Vvedenie v osobennosti odesskogo yazyka] Helsinki, available at: http://www.helsinki.fi/venaja/opiskelu/graduja/kabanen.pdf
5. *Mechkovskaya N. B.* (2006) Russian in Odessa: Yesterday, today, tomorrow [Russkiy yazyk v Odesse: vchera, segodnja, zavtra], Vol. 30, No. 2. pp. 263–281.
6. *Paljko M. L.* (2008) Intonation of the German coherent discourse in contrast to the Russian one [Intonatsiya nezavershennosti teksta v nemetskom yazyke v sopostavlenii s russkim], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2008" [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii "Dialog 2008"], Issue. 7 (14), pp. 416–419, available at: http://www.dialog-21.ru/digests/dialog2008/materials/html/65.htm
7. *Svetozarova N. D.* (2007) Concise Review of Phonetics of Yiddish as contrasted to Russian and German [Kratkiy ocherk fonetiki yazyka idish v sravnenii s russkim i nemetskim], Acta Linguistica Petropolitana [Linguistic issues of Saint Petersburg], Vol. 3, Section 1, pp. 224–239.
8. *Stepanov E. M.* (2004) The Russian Speech of Odessa [Rosiys'ke movlennya Odesy], Astroprint, Odessa.
9. *Yanko T.* (2008) Intonatsionnye strategii russkoj rechi v sopostavitel'nom aspekte [Intonational strategies of the Russian speech from a contrastive perspective]. Yazyki slavyanskih kul'tur, Moscow.
10. *Grice M., Baumann S., Benzmüller R.* (2005). German Intonation in Autosegmental-Metrical Phonology, Prosodic Typology: The Phonology of Intonation and Phrasing. Oxford University Press, available at: http://www.coli.uni-saarland.de/publikationen/softcopies/Grice:19xx:GIA.pdf
11. *Groennum N.* (1998) Fonetik og Fonologi. Almen og Dansk. [General and Danish Phonetics and Phonology] Akademisk Forlag, Copenhagen.
12. *Odé C.* (2002) An Outlook of describing and transcribing the Russian intonation in the spoken corpora [Perspektivy opisaniya i trankriptsii russkoy intonatsii v korpusakh zvuchashchikh tekstov], Problems and methodology of experimental phonetic investigations. Towards the 70th Anniversary of professor L. V. Bondarko [Problemy i metody eksperimental'no-foneticheskikh issledovanij. K 70-letiyu professora L. V. Bondarko], Saint Petersburg, pp. 209–214.
13. *Pierrehumbert J.* (1980) The Phonology and Phonetics of English Intonation, MIT PhD Dissertation, available at: http://faculty.wcas.northwestern.edu/~jbp/publications/Pierrehumbert_PhD.pdf

СОЧЕТАЕМОСТЬ ЧЕРЕЗ ПРИЗМУ КОРПУСОВ

Захаров В. П. (vz1311@yandex.ru)

Санкт-Петербургский государственный университет,
Санкт-Петербург, Россия

В статье рассматриваются устойчивые сочетания разного типа и показываются способы их количественной оценки. Описаны эксперименты, в ходе которых на материале корпусов русского языка и инструментов корпусной лингвистики показано, как с помощью корпусных методов можно расширить состав словарных статей в словарях устойчивых выражений и как можно количественно оценить употребительность и устойчивость словосочетаний в синхронии и диахронии.

Ключевые слова: устойчивые словосочетания, фразеологизмы, коллокации, корпусы текстов, меры ассоциации, диахронические исследования

SET PHRASES: A VIEW THROUGH CORPORA

Zakharov V. P. (vz1311@yandex.ru)

Saint-Petersburg State University, Saint-Petersburg, Russia

The study of word collocatibility is one of the main tasks of linguistics. Syntagmatic relations bind together language units being in direct contact with each other. The combinatory ability of language units, collocatibility, is one of the linguistic syntagmatic laws. This phenomenon is the main object of the phraseology and lexicography. The article deals with set phrases of different types from the point of view of their numerical evaluation. Corpus linguistics understand set phrases as statistically determined unities. This approach is the basic point of different automatic ways to extract idioms and collocations. The paper describes experiments which show how text corpora and corpus methods and tools such as association measures, word sketches, concordances can be used to expand the entries in existing dictionaries and how set phrases could be evaluated quantitatively. There are a small numbers of works on set phrases productivity during time periods because of small size of historical corpora. In this research examined set phrases usage was studied diachronically on the base of the big Google books Ngram Viewer Russian corpus counting billions of tokens. The study argues that diachronic productivity is best evaluated with a studying contexts. Used corpus tools enable to do it. Ultimately, it is shown and maintained that corpus linguistics methods and tools allow to create dictionaries of new type which have to include a larger amount of set phrases and collocations than before.

Key words: set phrases, idioms, collocations, collocation dictionaries, corpus, association measures, concordance, diachronical research

Введение

Один из популярных предметов в языкознании — это устойчивые словосочетания. Несмотря на пристальное внимание лингвистов к фразеологии и связанным темам, можно утверждать, что состояние фразеологии сегодня, на наш взгляд, неудовлетворительно. Фразеологический запас русского языка разбросан по разным лексикографическим изданиям, прежде всего это толковые и фразеологические словари, и ни один словарь не может считаться достаточно полным по охвату фразеологического лексикона. Предположительно, словник такого полного словаря должен насчитывать несколько сотен тысяч единиц. Также нетрудно убедиться, что статьи существующих фразеологических словарей неполны, они плохо структурированы, никак не привязаны к хронологии. В наши дни эту ситуацию можно существенно улучшить. Наличие корпусов текстов создает предпосылки для создания большого словаря сочетаемости, основанного на корпусах, с количественной параметризацией внутри.

Нужно отметить, что в корпусной лингвистике сложилась методология более широкого понимания фразеологии, и границы фразеологии здесь значительно расширены (или размыты) за счет новых подходов, общим для которых является понятие «статистической устойчивости». Может быть, самой знаменитой цитатой в корпусной лингвистике является высказывание Дж. Р. Фёрса «Вы поймете слово по его окружению» (“You shall know a word by the company it keeps”) [Firth 1957: 179]. Там же и тогда же им было введено вошедшее сегодня в широкий оборот понятие коллокации, базирующееся на статистических критериях. Об этом же писал И. А. Мельчук. «Устойчивость сочетания относительно данного элемента измеряется вероятностью, с которой данный элемент предсказывает совместное появление остальных элементов сочетания (в определенном порядке относительно предсказывающего элемента)» [Мельчук 1960: 73].

В корпусной лингвистике в основе методов вычисления силы синтагматической связи между элементами словосочетаний лежат частотные характеристики и структурно-синтаксические модели, на основе которых по формулам так называемых ассоциативных мер (мер ассоциации) вычисляется коэффициент силы связанности или, по-другому, уникальности словосочетания.

Данное исследование преследует цель показать, как можно улучшить словари сочетаемости корпусными методами. Была поставлена задача — на основе корпусных данных изучить «поведение» некоторых устойчивых словосочетаний, в том числе в течение длительного промежутка времени.

1. Материал и инструмент исследования

В качестве материала и инструмента исследования были использованы Национальный корпус русского языка (НКРЯ) (http://ruscorpora.ru), корпусы русских текстов ruTenTen 2011 и ruTenTen 2011 sample системы Sketch Engine (https://the.sketchengine.co.uk/), корпус русских текстов Araneum Russicum Maius из семейства псевдопараллельных корпусов Aranea Университета им. А. Коменского

в Братиславе (http://ucts.uniba.sk/) [Benko 2013], русский корпус системы Google books Ngram Viewer и, соответственно, их программные средства. Объем основного корпуса НКРЯ составляет 230 млн словоупотреблений, ruTenTen 2011 sample и русскоязычный Araneum насчитывают по 1,2 млрд токенов (около 1 млрд текстоформ), ruTenTen 2011 имеет объем более 18 млрд токенов (14,5 млрд текстоформ). Самый же большой из них — корпус русских книг Google books Ngram Viewer (https://books.google.com/ngrams). В настоящее время это наиболее мощный инструмент диахронических исследований. Эта система содержит корпусы размеченных текстов книг на 9 языках. Корпус книг на русском языке содержит 591 310 текстов общим объемом более 67 млрд словоупотреблений. Самые поздние публикации, включенные в корпус, относятся к 2008 году.

Основной лексической единицей (ЛЕ), с которой работает данная система, является N-грамма — последовательность от одной до пяти словоформ. Причем N-грамма, для того чтобы быть учтенной и обработанной, должна встретиться в корпусе не менее 40 раз. Для каждой заданной ЛЕ для заданного временного интервала строится единый график, по вертикальной оси которого откладывается относительная частота встречаемости заданных N-грамм в корпусе в данном году (частота, деленная на общее число словоупотреблений в корпусе за этот год), выраженная в процентах. На горизонтальной оси показаны годы, входящие в заданный временной интервал. *Каждая кривая графика маркируется цветом, в конце кривой указывается, какой* N-грамме (слову или словосочетанию) она соответствует (рис. 1).

При построении графиков изменения частоты употребления ЛЕ используется так называемое «сглаживание» (smoothing) При нулевом сглаживании в графике учитывается относительная частота встречаемости N-граммы за каждый год. Однако тенденция в динамике «поведения» слов более отчетливо прослеживается при скользящем усреднении данных. Если значение коэффициента сглаживания равно 3, то это означает, что для некоторого года к числу словоупотреблений искомого слова за этот год прибавляется число словоупотреблений его за три предыдущих года и три последующих и полученная сумма делится на семь. Относительное значение этой средней величины в процентах отражается на вертикальной оси.

Имеется тег «подстановочный знак» * (wildcard). Ввод его через пробел после N-граммы или до неё позволяет построить график встречаемости десяти наиболее частотных сочетаний данной N-граммы и слова, следующего за нею или ей предшествующего. Кроме построения графиков, система предоставляет ссылки к текстам, где встретились заданные ЛЕ. Как правило, это библиографические описания книг и фрагменты текстов с выделением в них заданных N-грамм. В некоторых случаях доступен полный текст книги в графическом формате. Более подробно о сервисе Google books Ngram Viewer см. [Захаров, Масевич 2014].

Похожий инструмент под названием «Графики» с 2012 г. работает и в составе НКРЯ. Функционально он подобен сервису Google books Ngram Viewer. Вход в этот сервис возможен как со страницы с результатами поиска по запросу к основному корпусу (ссылка *Распределение по годам*), так и из главного меню (ссылка *Графики*). При сходной идеологии, формулы подсчета относительной частоты и сглаживания в сервисах Национального корпуса и Google Ngram Viewer отличаются. Имеется возможность показать таблицы с абсолютными

и относительными частотами заданных ЛЕ за каждый год. Из таблиц по гиперссылкам возможен переход к просмотру примеров из корпуса.

2. Эксперименты

В качестве примеров устойчивых сочетаний для исследования были выбраны сочетания двух типов: 1) свободные сочетания с характерными определениями к слову «аплодисменты», выражающими, говоря в терминах теории «Смысл — Текст», функцию Magn; 2) фразеологизированные сочетания с глаголом «перебиваться» в значении «бедствовать».

2.1. «Аплодисменты»

Посмотрим, какие стандартные определения к слову «аплодисменты» зафиксированы в словарях. Новый Большой академический словарь приводит следующие сочетания: *бурные аплодисменты*, *гром аплодисментов* [БАС 2004]. Словарь сочетаемости слов русского языка дает: *громкие, продолжительные, долго не смолкающие, несмолкаемые, бурные, дружные, одобрительные, горячие, восторженные аплодисменты,* а также *сдержанные, скупые, редкие, жидкие* [Словарь сочетаемости 1983]. Неизвестно, что отражает порядок их следования.

В системе Sketch Engine имеется инструмент вычисления коллокаций по 7 мерам ассоциации. В одном из режимов мы получили список из 36 прилагательных, из которых 19 можно считать функцией Magn от слова «аплодисменты». Вот список этих прилагательных, упорядоченный по алфавиту: *бешеный, бурный, дружный, восторженный, всеобщий, горячий, громкий, несмолкающий, громовой, громогласный, долгий, дружный, неистовый, нескончаемый, несмолкаемый, несмолкающий, оглушительный, продолжительный, шумный.*

Традиционные словари ничего не говорят о частоте употребления словарных единиц. Эти данные можно получить из корпусов. Например, поиск в НКРЯ (интервал 3 слова вправо) в данном случае дает следующие цифры: *бурные аплодисменты* 337 словоупотреблений, *продолжительные* 125. *дружные* 81, *громкие* 47, *оглушительные* 22, в то время как *несмолкающие* всего одно.

При этом важно понимать, в каком корпусе мы ищем, что и как ищем. Так, при поиске словосочетания *бурные аплодисменты* в НКРЯ в интервале в одно слово вправо оно было найдено в 279 контекстах, в то время как поиск в интервале в 3 слова дает нам 337 контекстов. Подавляющая часть прироста обеспечивается релевантным сочетанием *бурные и продолжительные аплодисменты*. Поиск же по сочетанию *дружные аплодисменты* в том же интервале выдает нам не совсем корректные сочетания *дружный смех и аплодисменты*, *дружный хохот и аплодисменты* и др. То есть, число 81 для *дружных аплодисментов* несколько завышено. Главное, полученные цифры не нужно абсолютизировать, важны их относительные величины.

Иногда, задав нестандартный режим поиска, можно получить дополнительно интересные результаты. Например, ни один из наших корпусов не дал

к *аплодисментам* коллоката *жесткий*. Однако поиск в интервале с отключенным согласованием между этими словами дает фразу *«По жесткому звуку аплодисментов чувствовалось…»*, где появляется это определение.

Выявление коллокаций по формулам мер ассоциации учитывает не только частоту совместной встречаемости, но и частость или редкость каждого элемента, тем самым мы вычисляем именно силу связи между элементами сочетания. Результат можно упорядочить или по частоте, или по значению меры ассоциации (табл. 1). И мы видим, что *продолжительные аплодисменты* по силе связи (мера salience) оказались лишь на 7-м месте.

Таблица 1. Примеры сочетаний «прилагательное + *аплодисменты*» в корпусе ruTenTen 2011, упорядоченных по мере ассоциации salience

№ п/п	Словосочетание	Частота в корпусе	Мера salience
1.	бурные аплодисменты	13 372	10,25
2.	дружные аплодисменты	2 051	8,42
3.	оглушительные аплодисменты	656	8,29
4.	громкие аплодисменты	3 711	8,22
5.	восторженные аплодисменты	899	8,17
6.	одобрительные аплодисменты	388	8,05
7.	продолжительные аплодисменты	2 495	7,80
8.	несмолкающие аплодисменты	211	7,62

Результаты поиска сочетаний со словом *аплодисменты*, полученные разработчиками Генерального Интернет-корпуса русского языка (ГИКРЯ) [Беликов и др. 2013] на их корпусе (подкорпус «Журнальный зал», объем 313 млн словоупотреблений) и любезно предоставленные автору, дают несколько другую картину, а именно: *бурные* 194, *продолжительные* 72, в т. ч. *бурные (и) продолжительные* 33, *громкие* 39, *дружные* 26, *шумные* 17, *восторженные* 15, *долгие* 14, *горячие* 13, *оглушительные* 12, *несмолкающие* 12. Это еще раз говорит о том, что цифры не нужно абсолютизировать и нужно учитывать, на каком корпусе они получены. При этом для оценки распространенности соответствующих единиц в корпусе (а тем самым в какой-то степени и в языке) нужно опираться не на абсолютные частоты, а на относительные (ipm). Проиллюстрируем это на маленьком примере (табл. 2).

Таблица 2. Сравнение частот сочетаний в разных корпусах

	Частота в корпусе			ipm		
Словосочетание	НКРЯ	ruTenTen	ГИКРЯ	НКРЯ	ruTenTen	ГИКРЯ
дружные аплодисменты	81	2 051	26	**0,350**	0,140	0,080
громкие аплодисменты	47	3 711	39	0,200	**0,260**	0,120
несмолкающие аплодисменты	1	211	12	0,004	0,015	**0,038**

Как видим, разные корпусы по-разному оценивают вес соответствующих сочетаний в языке, а фактически, в подъязыке, который представлен корпусом. Отдельная задача — попытаться эту разницу понять и объяснить, с тем чтобы какие-то особенности корпуса (преобладание какой-то тематики или типа текстов, возможное наличие дублетов и т.п.) не переносить на язык в целом.

Тем не менее, создавая словарь устойчивых сочетаний на основе корпусов, мы имеем возможность выстроить их по частоте употребления или по силе «спаянности», оговорив, на каком материале этот словарь создается. Более того, по-видимому, иногда полезно опираться на усредненные характеристики, полученные на разных корпусах.

Данные, полученные на синхронных корпусах, ничего не говорят о продуктивности ЛЕ в разные промежутки времени. Чтобы увидеть их использование на протяжении длительного периода, построим графики распределения частоты употребления наших сочетаний в сервисах Google books Ngram Viewer и «Графики» НКРЯ в текстах двух последних столетий (рис. 1).

Рис. 1. Кривые встречаемости биграмм со вторым словом «аплодисменты» в корпусе Google books Ngram Viewer (сглаживание 3)

Как мы уже видели, «лидируют» *бурные аплодисменты*, а на втором месте — *продолжительные*. Но это суммарные данные по всему корпусу. На графиках же мы видим особенности распределения частот употребления этих сочетаний во времени — см. пики на рубеже 1940-х, 1960-х и 1980-х годов. И если во второй половине 1930-х «верх берут» *бурные*, то в конце 70-х — начале 80-х преобладают *продолжительные*. Следует отметить, что в широкое употребление все эти сочетания вошли только в XX веке. А в конце века их частотность резко упала. Это видно и из сервиса Google, и сервиса НКРЯ. Однако сервис НКРЯ в этом и в других случаях дает картину мало репрезентативную по причине недостаточности данных. Анализируемые словосочетания представлены в корпусе в малых количествах и не в каждом году. Так, *громкие аплодисменты* встретились в основном корпусе НКРЯ по одному разу в 1885, 1906, 1908, 1910, 1925, 1939–40, 1959, 1963, 1998–2000, 2003 гг. и два — в 2001 г. Этого явно мало. Поэтому мы опираемся на графики системы Google.

Если же задать сглаживание, равное нулю, то можно определить пик использования того или другого сочетания в каждом году (точнее, в текстах

данного года; напомним, в корпусе Google books Ngram Viewer это книги). Для *продолжительных аплодисментов* «рекордным» оказался 1981 г.

Проанализируем также атрибутивное отношение, выраженное в форме «существительное в им. пад. + *аплодисменты* в род. пад.». Все словари согласно приводят следующие коллокаты для выражений этого типа: *шквал, гром, буря, грохот, взрыв*. Остается, однако, неясной частотность их употребления. Данные поиска в корпусах ruTenTen 2011 sample и Araneum Russicum Maius и график (рис. 2) показывают явное преимущество коллоката *гром*.

Рис. 2. Кривые встречаемости биграмм с существительным и со вторым словом «аплодисменты» в родительном падеже в корпусе Google books Ngram Viewer (сглаживание 3)

На втором месте — *буря*. Но мы видим, что начиная с 1980-х годов *шквал аплодисментов* идет вверх и устойчиво обгоняет *бурю*. А если обратиться к корпусу ГИКРЯ, отражающему современное состояние языка, то в подкорпусах «Живой журнал» и «В контакте» *шквал* уже обошел и *гром*, что говорит о том, что «живой» язык, видимо, предпочитает последнее сочетание.

Поиск в основном корпусе НКРЯ дает 118 вхождений для *гром аплодисментов* (165 с учетом словоизменения, но следует помнить, что графический сервис работает со словоформами), 33 вхождения для *шквал аплодисментов*, 15 для *буря аплодисментов*, только одно для *взрыв аплодисментов* и ни одного для *грохота*.

Однако можем ли мы полностью доверяться конкретному корпусу? Последние два сочетания в 230-миллионном НКРЯ фактически не представлены, зато в соизмеримом с НКРЯ по объему подкорпусе ГИКРЯ («Журнальный зал») нашлось 25 *взрывов* и 10 *грохотов*. Этот и подобные факты требуют своего объяснения, чтобы мы могли опираться на получаемые результаты либо иногда их отбрасывать. Например, частотное слово или словосочетание может иметь источником этой частоты его «взрывообразную» представленность всего лишь в нескольких текстах в коротком промежутке времени и не быть характерным для языка в целом. Для минимизации таких «всплесков» в лингвистике существуют специальные меры, учитывающие равномерность появления слова в корпусе (коэффициент Жуйана, Average Reduced Frequency и др.).

И последнее: следует учитывать лексико-синтаксическое варьирование исследуемых сочетаний. Так, в корпусе Araneum Russicum Maius корпусный менеджер насчитал 14 *взрывов аплодисментов*, в то время как сочетание *взорваться аплодисментами* в разных формах глагола встретилось 42 раза, что говорит о том, что набор конструкций, выражающих функцию Magn от слова «аплодисменты», должен быть расширен.

2.2. «Перебиваться»

Многие фразеологизмы и устойчивые сочетания имеют лексико-синтаксические варианты, когда либо меняется лексическое наполнение в рамках некоторой структурной формулы, либо при том же наполнении меняется формула. Например, «кошки скребут». Но где? Словари сообщают, что *на душе* и *на сердце*. А где чаще? По данным корпуса Google books Ngram Viewer выясняется, что чаще *кошки скребут на душе* и больше всего они скребли в годы на переломе 1980–90-х гг.

Наверное, не будет ошибкой утверждение, что фразеологизмов с лексико-синтаксическими ввариациями большинство. Примеры их можно множить и множить: *беречь (хранить) как зеницу ока; беречь пуще глаза; мерить одной мерой (меркой), мерить на одну меру (мерку); ест за троих, есть в три горла; драть (сдирать/содрать) шкуру (три, две шкуры); драть (сдирать/содрать) по три (две) шкуры;хоть в землю заройся, хоть из-под земли достань; брать/взять (забирать/забрать) в [свои] руки, прибирать/прибрать к рукам; сталкивать/столкнуться лицом к лицу, носом к носу, нос в нос, лоб в лоб* [Бирих и др. 1997]. Такие вариативные сочетания в словарях описаны, естественно, менее полно по сравнению с лексикализованными фраземами.

Рассмотрим этот тип вариаций более подробно на примере сочетания глагода «перебиваться» с предложной конструкцией «с … на …». Традиционно словари дают сочетания *перебиваться с хлеба на квас, перебиваться с хлеба на воду*. Кроме того, приводятся синонимические конструкции *с куска на кусок, с гроша на копейку, с пуговки на петельку* [Бирих и др. 1997: 15].

Посмотрим, что нам дополнительно дают корпусы. Поиск в указанных корпусах добавляет к вышеприведенному списку еще немалую толику, а именно: *с хлеба на воду,* с *хлеба на кофе, с гроша на грош, с копейки на копейку, с рубля на рубль, с хлеба на квас, с воды на квас, с воды на хлеб, с хлеба на картошку, с петельки на пуговку, со дня на день, с весны на весну, с работы на работу.* Есть и более экзотические: *с селедки на вермишель, с "Российского" на "Докторскую"*. А к вершинам народного языкового творчества можно отнести вот это: «И это беспроводной интернет!.. У Мегафона четкий прием — 3894 Кбит/с. Мне, молившемуся на dial-up, который *перебивался с 22 на 40 килобит в секунду*, это кажется чем-то фантастическим».

Вышеприведенные примеры в основной массе получены из корпусов, созданных по технологии Wacky, то есть на основе текстов из веба, и нередко они отражают языковое творчество, но не узус, подчеркивая лишь продуктивность

конструкции «перебиваться с … на …». Однако наличие больших корпусов позволяет выявлять значительное число кандидатов на вхождение в словарь в качестве устойчивых сочетаний, а статистические данные дают возможность оценить распространенность того или другого сочетания.

Из графика на рис. 3 видно, что наиболее частые устойчивые сочетания с глаголом «перебиваться» это те, которые приводятся в фразеологических словарях, и что активно в языковой обиход они вошли только в начале XX века. Зато сочетание *перебиваться со дня на день* широко использовалось в XIX веке.

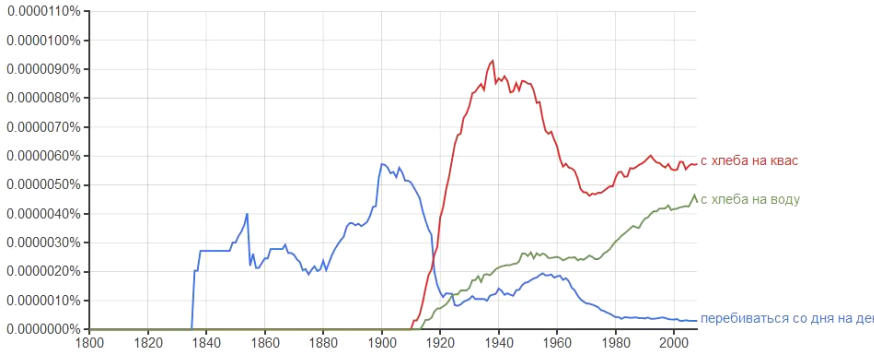


Рис. 3. Кривые встречаемости выражений с глаголом «перебиваться» в корпусе Google books Ngram Viewer (сглаживание 3)

И здесь еще раз следует подчеркнуть, что, интерпретируя корпусные данные, мы должны хорошо понимать, что собой представляет тот или другой корпус и как эти данные получены. Например, данные анализа по корпусу ГИКРЯ, предоставленные автору его разработчиками, показывают, что если в корпусе книг Google сочетание *с хлеба на квас* встречается чаще, чем *с хлеба на воду*, то во всех трех подкорпусах ГИКРЯ картина диаметрально противоположная: в сумме 269 употреблений *с хлеба на воду* против 97 *с хлеба на квас*. То же соотношение демонстрирует и корпус ruTenTen 2011 (787 против 370). Все это позволяет говорить о различии в использовании этих выражений между книжным языком и современным «спонтанным».

3. Опыт корпусного исследования

Проведем небольшое исследование на тему, как выглядит в корпусах сочетаемость слова «мастер». Сочетаемость слов определяется различными факторами: лексическими, грамматическими, семантическими, стилистическими. Все они влияют на норму и на узус. Можно утверждать, что узус — один из определяющих факторов при составлении словарей. И один из подходов к изучению узуса заключается в выявлении статистических закономерностей на корпусах текстов.

Слово «мастер» в русском языке является довольно частотным. Его ipm по электронной версии Частотного словаря современного русского языка (на материалах Национального корпуса русского языка) О. Н. Ляшевской и С. А. Шарова (http://dict.ruslang.ru/freq.php) равняется 100,8, в корпусе ruTenTen 2011 он равен 96,0. В словарях его сочетаемостные свойства никак особенно не описываются.

Рассмотрим его «поведение» на материале вышеупомянутых корпусов. При изучении конкордансов с этим словом обращаешь внимание на сочетание «каких-то дел мастер». И таких контекстов, кроме обычно приводимого в словарях *золотых дел мастер*, в корпусах находится достаточно много (329 в НКРЯ и 4015 в ruTenTen 2011). Если объединить все определения к словоформе *дел*, то их будет 252:

абордажных, автомобильных, аккордеонных, алмазных, багетных, балаганных, банкетных, банных, барабанных, баррикадных, бархатного, берестяных, библиотечных, бриллиантовых, броневых, бронзовых, бронзовых, бронно-кольчужных, булочных, бумажных, буровых, бытовых, взрывных, винных, витражных, водочных, воровских, выборных, вывесочных, газетных, газовых, гаражных, гармонных, гитарных, глазных, гламурных, глиняных, гончарных, горных, городовых, грильных, гробовых, дамских, дверных, деревянных, деспотических, дипломатических, добрых, домашних, дорожных, железных, жестяных, живописного, журнальных, закулисных, замочных, заплатных, заплечных, запретных, здоровых, земельных, зеркальных, золотых, игрушечных, изразцовых, именных, иностранных, искусных, кабельных, кальянных, каменных, каминных, каретных, карманных, картежных, кирпичных, ключных, книжных, ковровых, кожаных, кожевенных, колбасных, колесных, колодезных, колокольных, колыбельных, кольчужных, комедийных, компьютерных, конкурсных, конфетных, коньячных, копировальных, корабельных, корабельных, костяных, кофейных, красочных, крепостных, крепостных, кроватных, кровельных, кровопийственных, кузнечных, кузовных, кукольных, кулачных, кулинарных, кухонных, ледяных, лепных, литейных, литературных, лодочных, любовных, макетных, малярно-живописных, малярных, машинных, мебельных, мебельных, медных, мироедских, мозаичного, мозольных, молочных, монетных, мостовых, музыкальных, музыкальных, мусийных, мясных, ножевых, обувных, огненных, оконных, оловянных, оптических, органных, оружейных, открыточных, палатных, палаточных, палаческих, памятных, парусных, переговорных, переплетных, персонных, перспективных, перчаточных, песочных, печатных, печных, плотницких, погребальных, поддельных, подкопных, подручных, подъемных, позолотных, половых, помойных, портновских, портных, портняжных, похоронных, почтовых, преоспективного, прикладных, пробирных, прохвостных, пушечных, пыточных, пытошных, ракетных, резных, рекламных, ресторанных, ритуальных, розыскных, ручных, рыбных, садовых, самолетных, сапожных, сателлитных, седельных, селедочных, сердечных, серебряных, сих, скрипичных, сладких, слесарных, словесных, социальных, ссудных, стегательных, стеклянных, стекольных, стекольных, столярных, страховых, строительных, суконных, сусалнаго, сценических, сыскных, табачных, табуреточных, тайных,

телевизионных, ткацких, токарных, топорных, трубных, угольных, ударных, фальшивых, фейерверкского, фершельных, фискальных, фонтанных, фотографических, фотошопных, хлебных, ходульных, холодильных, хрустальных, цветочных, ценинных, цеховых, циркульных, чайных, часовых, чеканного, чемоданных, черепаховых, чернильных, шапочных, шашлычных, швейных, шлифовальных, шлюзных, шляпных, шляпочных, шоколадных, ювелирных, янтарных.

Анализ показывает, что за исключением немногих, окказиональных или оценочных, все они относятся к какому-либо ремеслу. Приведем 25 наиболее частотных сочетаний, полученных на корпусе ruTenTen (табл. 3).

Таблица 3. Наиболее частотные прилагательные в сочетаниях типа «таких-то дел мастер», упорядоченные по частоте совместной встречаемости

Ранг по частоте	Слово	Частота
1.	золотых	519
2.	часовых	261
3.	серебряных	179
4.	заплечных	112
5.	каменных	111
6.	кукольных	92
7.	гробовых	73
8.	пыточных	58
9.	витражных	57
10.	добрых	49
11.	оружейных	46
12.	кузнечных	46
13.	кузнечных	46

Ранг по частоте	Слово	Частота
14.	колокольных	44
15.	ювелирных	40
16.	шляпных	39
17.	чемоданных	38
18.	обувных	38
19.	похоронных	37
20.	мебельных	37
21.	деревянных	36
22.	сапожных	33
23.	печатных	31
24.	скрипичных	30
25.	столярных	27

Высокая величина частоты совместной встречаемости, казалось бы, говорит об устойчивости данного сочетания. Однако этой характеристики недостаточно, чтобы говорить о предпочтительной сочетаемости одного слова с другим. Сочетание с невысокой частотой совместной встречаемости может представлять собой неделимое единство. Имеется целый ряд статистических мер (меры ассоциации, англ. association measures), вычисляющих силу «спаянности» сочетаний. Значения мер ассоциации можно считать показателями силы синтагматической связи между элементами словосочетаний. Приведем те же 25 сочетаний с подсчитанными значениями нескольких мер ассоциации (табл. 4).

Мы видим, что коллокации «перестроились» и что разные меры по-разному оценивают силу синтагматической связи. И не всегда бóльшую силу связи получают наиболее частые сочетания, например, сочетание *витражных дел мастер*, всего лишь девятое по частоте (табл. 4), оказывается первым по рангу меры MI3 и, по-видимому. с бóльшим основанием может быть включено в словарь в качестве (или как пример) устойчивого словосочетания.

Таблица 4. Наиболее частотные прилагательные в сочетаниях типа «таких-то дел мастер», упорядоченные по значению меры ассоциации MI3

Ранг по частоте	Слово	Частота сочетания	Ранг по MI3	MI3	log likelihood	log Dice	MI. log_f
9.	витражных	57	*1.*	33,096	1472,831	8,651	85,745
1.	золотых	519	*2.*	32,046	9180,223	6,822	87,557
4.	заплечных	112	*3.*	31,295	2132,625	9,131	82,657
7.	гробовых	73	*4.*	30,744	1421,884	8,719	77,587
2.	часовых	261	*5.*	30,188	4034,095	7,368	78,951
17.	чемоданных	38	*6.*	29,400	822,664	7,993	68,098
13.	кузнечных	46	*7.*	28,675	520,601	7,321	60,985
20.	мебельных	37	*8.*	28,365	594,114	7,550	61,788
16.	шляпных	39	*9.*	28,330	771,052	7,909	64,324
3.	серебряных	179	*10.*	28,301	2676,989	6,488	69,416
8.	пыточных	58	*11.*	28,261	971,907	8,109	66,014
6.	кукольных	92	*12.*	27,395	1239,558	7,604	64,266
14.	колокольных	44	*13.*	27,085	799,089	7,754	60,775
23.	печатных	31	*14.*	27,023	584,485	7,532	58,078
22.	сапожных	33	*15.*	26,760	556,710	7,464	56,926
24.	скрипичных	30	*16.*	25,642	534,014	7,264	53,529
5.	каменных	111	*17.*	25,584	1529,471	5,062	56,727
10.	добрых	49	*18.*	25,089	650,132	6,918	53,147
11.	оружейных	46	*19.*	25,050	719,464	6,724	53,500
21.	деревянных	36	*20.*	25,019	414,504	7,010	49,856
12.	кузнечных	46	*21.*	24,712	500,269	6,942	50,463
19.	похоронных	37	*22.*	24,203	510,628	6,642	49,124
18.	обувных	38	*23.*	23,816	509,175	6,364	47,956
25.	столярных	27	*24.*	22,038	344,005	5,640	40,968
15.	ювелирных	40	*25.*	20,973	423,087	3,777	38,368

В лексикографических изданиях следует, вероятно, также указать, что в этих сочетаниях слово *дело* почти всегда стоит во множественном числе (из 329 сочетаний в НКРЯ только в 17 *дело* в единственном числе). Также нужно упомянуть, что в этого рода сочетаниях имеет место именно такой порядок слов. Конечно, не будет большой ошибкой сказать, *мастер кузнечных дел*, но говорят ли так? Проверка на большом «живом» материале дает нам ответ: говорят, но редко. Вот что показывает корпус ruTenTen 2011: *мастер золотых дел* 12 сочетаний против 579 с *мастером* в постпозиции, *мастер серебряных дел* 12 против 179, *мастер гробовых дел* 2 и 73, *мастер пыточных дел* 1 и 58.

Стоит упомянуть в словарях и характерное сочетание *сих дел мастер*. В НКРЯ оно встречается 4 раза, в корпусе ruTenTen 2011 12, но и во всем вебе в Яндексе — всего 722 раза, причем это все дубли, а разных цитат немногих более 12.

Наиболее часто представлено высказывание Л. Троцкого о В. Ленине — о склоке, «которую разжигает *сих дел мастер* Ленин, этот профессиональный эксплуататор всякой отсталости в русском рабочем движении». Изредка это сочетание встречается в литературе, в частности, у Н. К. Михайловского, А. П. Чехова, Н. А. Тэффи, С. В. Максимова, Н. Е. Врангеля, И. Н. Потапенко, В. Ф. Пановой, причем нередко в кавычках. На самом деле же это сочетание того же профессионально-ремесленного происхождения, что и вышеприведенные выражения. Вот откуда это пошло, читаем: *«Как в Киеве я смеялся, смотря на вывеску, на которой были изображены самовар, мельница и ножницы и подписано: «сих дел мастер»…»*.

Также для анализа сочетаемости представляют интерес предложные сочетания со словом *мастер* (предлоги «по», «на», «с», «в», «от»), где оно выступает как «хозяин» в структуре зависимостей, то есть управляет предлогом, связывающим его со знаменательным словом.

Рассмотрим здесь сочетания только с предлогом «на». Очень часто это синтаксема (отношение между хозяином, предлогом и слугой), которую Г. А. Золотова определяет как потенсив — «синтаксема от отвлеченных имен, обозначающих потенциальное действие при словах модальной семантики (глаголах, именах, прилагательных). С личными именами (мастер, мастак, охотник) потенсив образует сочетание, представляющее собой модальную и экспрессивно-оценочную модификацию предикативной характеристики лица» [Золотова 1988: 197]. Отметим, что для таких сочетаний имеется синонимичная конструкция с глаголом (*мастер на шутки — мастер шутить*).

Какие же «процессные существительные» встречаются в корпусах в сочетании с *мастером*? В корпусе ruTenTen 2011 находится 34 таких сочетания со словом *дело*, причем *дело* всегда стоит во множественном числе и с определением. Частые определения к *делам эти* и *такие*, кроме того, встретились *темные, плохие, маленькие* и *пытошные*. В числе других «потенциальных действий» для слова *мастер* были найдены: *штуки, штучки, шутки, проделки, операции, авантюры, интриги, трюки, разговоры, анекдоты, выдумки, флешмобы, проказы, хитрости, слова*. Почти всегда с определениями, среди которых преобладают *такие, всякие, подобные, разные, всевозможные*, также встречаются *сходственные, веселые, подлые, хаккерские, жестокие, недобрые*. В 15 случаях следом за предлогом идут сочетания *всякого, такого, разного, различного рода*.

Встречаются в корпусах и фразеологизированные выражения:

- *мастер на все руки* (78 раз в НКРЯ, 2910 раз в ruTenTen 2011);
- *мастер с большой буквы* (4 вхождения в НКРЯ и 470 в ruTenTen 2011, еще 7 вхождений в ruTenTen с другими определениями к *букве*: *самая большая, высокая, огромная, маленькая, та самая, вышитая*);
- *дело мастера боится* (30 вхождений в НКРЯ, 260 в ruTenTen 2011, с определениями *всякое, любое, ночное дело*).

Интересны сочетания *мастера* с выражениями на *свой лад, вкус, глаз* — в этом случае *мастеру* всегда предшествуют определения *всякий, всяк, каждый*.

Есть и другие сочетания для слова *мастер* и с другими предлогами, но на этом мы здесь остановимся.

4. Заключение и выводы

Сегодня русский язык переживает период быстрого обновления своего состава. Не избежали этого и устойчивые сочетания. На периферии языка оказываются сочетания, отражающие некоторые стороны социальной жизни дореволюционного общества (мир чиновничества, картежные игры и др.). Зато повышенную частотность получают сочетания из области науки, техники, спорта. Чтобы увидеть все эти изменения, нужны большие корпусы, особенно для фразеологии, учитывая сравнительно низкую частоту употребления фразеологизмов в текстах. И сейчас такие корпусы начинают появляться.

В то же время необходимо, чтобы корпусная лингвистика развивала свои средства. Так, система Google Books Ngram Viewer предоставляет большие возможности для историко-культурных и лингвистических исследований. Однако в текстах корпуса встречается много ошибок распознавания. Поиск заданных лексических единиц ведется по словоформам, а не по леммам. Корпус построен исключительно на книгах и тем самым не сбалансирован. По-видимому, целесообразно было бы провести основательное исследование с применением методов статистической обработки данных, чтобы понять, как эти и другие проблемы влияют на достоверность получаемых результатов. Все это относится и к сервису НКРЯ «Графики» (главный недостаток малый для полноценных диахронических исследований объем корпуса).

Проведенное исследование показало, что корпусы и инструментарий корпусной лингвистики позволяют выявить и существенно расширить лексический фонд устойчивых словосочетаний разного типа и особенности их бытования. Основываясь на корпусах, лингвисты имеют возможность создавать словари и учебники нового типа, где сочетаемость будет представлена неизмеримо шире, чем до сих пор. В качестве примера такого словаря можно привести словарь «КроссЛексика», в котором словосочетания составляют самую важную и самую объемную его часть (2,26 млн словосочетаний) [Большаков 2009]. При этом они должны иметь количественные характеристики как силы устойчивости в синхронии, так и историю их употребления в диахронии.

В ходе исследования мы также неоднократно убеждались, что для того, чтобы можно было делать достоверные выводы на основе корпусных данных, следует хорошо представлять себе недостатки и ограничения тех инструментов, которыми мы пользуемся.

Литература

1. *БАС* — Большой академический словарь русского языка. Том 1. М.—СПб.: Наука, 2004.
2. *Беликов В. И., Копылов Н. Ю., Пиперски А. Ч., Селегей В. П., Шаров С. А.* Корпус как язык: от масштабируемости к дифференциальной полноте // Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной Международной конференции «Диалог» (Бекасово, 29 мая — 2 июня 2013 г.). Вып. 12 (19). М.: Изд-во РГГУ, 2013. С. 84–95.

3. *Бирих А. К., Мокиенко В. М., Степанова Л. И.* Словарь фразеологических синонимов русского языка. Ростов-на-Дону, 1997.
4. *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.: Изд-во РГГУ, 2009. С. 45–50.
5. *Захаров В. П., Масевич А. Ц.* Диахронические исследования на основе корпуса русских текстов Google books Ngram Viewer // Структурная и прикладная лингвистика. Вып. 10. СПб.: Изд-во С.-Петербю ун-та, 2014. С. 303–327.
6. *Золотова Г. А.* Синтаксический словарь. М.: Наука, 1988.
7. *Мельчук И. А.* О терминах «устойчивость» и «идиоматичность» // Вопросы языкознания. 1960, № 4. С. 73–80.
8. *Словарь сочетаемости слов русского языка* / Институт русского языка им. А. С. Пушкина; Под ред. П. Н. Денисова, В. В. Морковкина. — 2-е изд., испр. — М.: Русский язык, 1983.
9. *Benko V. Aranea: Yet Another Family of* (Comparable) Web Corpora. In: Petr Sojka, Aleš Horák, Ivan Kopeček and Karel Pala (Eds.): Text, Speech and Dialogue. 17th International Conference, TSD 2014, Brno, Czech Republic, September 8–12, 2014. Proceedings. LNCS 8655. Springer International Publishing Switzerland, 2014, pp. 257–264. ISBN: 978-3-319-10815-5.
10. *Firth, J. R.* 1957. A synopsis of linguistic theory 1930–1955. In: F. Palmer (Ed.), Selected Papers of J. R. Firth 1952–1959. London: Longman, pp. 168–205.

References

1. *BAS* (2004), Great Academic Dictionary of the Russian Language, [Bol'shoy akademicheskiy slovar' russkogo yazyka], vol. 1, MoscowSaint-Petersburg, Nauka.
2. *Belikov V., Kopylov N., Piperski A., Selegey V., Sharoff S.* (2013), Corpus as language: from scalability to register variation, [Korpus kak yazyk: ot masshtabiruyemosti k differentsial'noy polnote], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2013" [Komp'juternaja lingvistika i intellektual'nye tekhnologii: po materialam ezhegodnoy mezhdunarodnoj konferentsii "Dialog 2013"], vol. 12 (19), Moscow, RGGU, pp. 84–95.
3. *Benko V.* (2014), Aranea: Yet Another Family of (Comparable) Web Corpora, In: Petr Sojka, Aleš Horák, Ivan Kopeček and Karel Pala (Eds.): Text, Speech and Dialogue. 17th International Conference, TSD 2014, Brno, Czech Republic, September 8–12, 2014. Proceedings. LNCS 8655. Springer International Publishing Switzerland, pp. 257–264, ISBN: 978-3-319-10815-5.
4. *Birikh A. K., Mokiyenko V. M., Stepanova L. I.* (1997), Dictionary of phraseological synonyms of the Russian language, [Slovar' frazeologicheskikh sinonimov russkogo yazyka], Rostov-on-Don.
5. *Bolshakov I. A.* (2009), CrossLexica: a large electronic dictionary of collocations and semantic links between Russian words, [KrossLexika — bol'shoy

elektronnyy slovar' sochetaniy i smyslovykh svyazey russkikh slov], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2009" [Komp'juternaja lingvistika i intellektual'nye tekhnologii: Trudy mezhdunarodnoj konferentsii "Dialog 2009"], vol. 8 (15), Moscow, RGGU, pp. 45–50.

6. *Collocability Dictionary of Russian Language Words* (1983), [Slovar' sochetayemosti slov russkogo yazyka], P. N. Denisov, V. V. Morkovkin (eds.), Moscow, Russkiy yazyk.
7. *Firth, J. R.* (1957), A synopsis of linguistic theory 1930–1955, In: F. Palmer (Ed.), Selected Papers of J. R. Firth 1952–1959, London, Longman, pp. 168–205.
8. *Melčuk I. A.* (1960), About the terms steadiness and idiomaticity, [O terminakh ‚ustoyvhivost" i ‚idiomatichnost"], Questions of Linguistics, [Voprosy yazykoznaniya], 1960, No. 4, pp. 73–80.
9. *Zakharov V. P., Masevich A. Ts.* (2014), Diachronic researches on the base of the Russian Google books Ngram Viewer text corpus [Diakhronicheskiye issledovaniya na osnove korpusa russkikh tekstov Google books Ngram Viewer], Srtuctural and Applied Linguistics [Strukturnaya i prikladnaya lingvistika], vol. 10, Saint-Petersburg, pp. 303–327.
10. *Zolotova G. A.* (1988), Syntactic dictionary [Sintaksicheskiy slovar'], Moscow, Nauka.

ЛИНГВОСПЕЦИФИЧНЫЕ ЕДИНИЦЫ РУССКОГО ЯЗЫКА В СВЕТЕ КОНТРАСТИВНОГО КОРПУСНОГО АНАЛИЗА[1]

Зализняк Анна А. (anna.zalizniak@gmail.com)

Институт языкознания РАН, Москва

Ключевые слова: русский язык, семантика, лингвоспецифичные единицы, параллельный корпус, контрастивный корпусный анализ, лексическая база данных

RUSSIAN LANGUAGE-SPECIFIC WORDS AS AN OBJECT OF CONTRASTIVE CORPUS ANALYSIS

Zalizniak Anna A. (anna.zalizniak@gmail.com)

IL RAS, Moscow, Russia

The paper summarizes methodological principles and some preliminary results of a project "Contrastive corpus-based study of the specific features of Russian semantic system" currently conducted by a research group on the basis of an aligned Russian-French and French-Russian parallel corpus. The purpose of the research project is to verify, by means of contrastive corpus-based analysis, a number of hypotheses concerning Russian "language-specific" words formulated in the course of previous investigations. We assume that translation equivalents of a language unit in another language can be considered as a source of information about the semantics of the latter. Such approach is particularly efficient in case of language-specific words that usually do not have full-fledged equivalents in other languages. Indeed, there are at least three possible types of mismatch: a certain semantic component is lacking in the translation equivalent; the translation equivalent includes an additional component, which is absent in the original unit; a certain semantic component is rendered by supplementary means. Each of these types of mismatch provides us with linguistic information that contributes to clarify the semantics of the source language unit and thus to verify the hypothesis of its language-specific status.

Key words: Russian language, semantics, language-specific words, parallel corpus, contrastive corpus analysis, lexical database

[1] Статья написана при финансовой поддержке РФФИ, грант № 13-06-00403.

На протяжении последних двух десятилетий лет мы с моими коллегами ведем исследования русской семантики в направлении, которое в какой-то момент получило название «реконструкция русской языковой картины мира». Результаты работ этого направления опубликованы в монографиях [Зализняк, Левонтина, Шмелев 2005, 2012], а также в ряде статей. За последнее десятилетие исследование «русской языковой картины мира» вошло в круг интересов российской когнитивной лингвистики и необычайно расширило свои рамки; одновременно укрепилась и критическая линия, подвергающая сомнению существование связи между национальным языком и менталитетом соответствующей языковой общности, как и само существование вышеозначенного менталитета. Наша позиция в этой дискуссии достаточно подробно изложена в работах [Зализняк 2013, 2014, Шмелев 2014], поэтому здесь я не буду воспроизводить аргументы в защиту — в конечном счете, гипотезы Сепира-Уорфа, — на которой базируются и которую подтверждают наши исследования. Мне хотелось бы перенести фокус внимания на более верифицируемую часть проблемы, а именно, на само понятие «лингвоспецифичных слов»: с одной стороны, уточнить его содержание, с другой стороны, предложить некоторую операциональную процедуру выявления принадлежащей к этой категории единиц на базе параллельных корпусов, т. е. при помощи анализа межъязыковых соответствий.

Именно такая задача ставится в проекте «Контрастивное корпусное исследование специфических черт семантической системы русского языка». Исследование проводится на базе параллельного русско-французского корпуса, который был создан, в значительной степени, силами коллектива данного проекта в ходе предшествующих этапов работы (см. наши публикации [Бунтман и др. 2014, Kruzhkov et al. 2014]) и в настоящий момент входит в качестве подкорпуса в Национальный корпус русского языка. На основе этого параллельного корпуса в ходе выполнения данного проекта была создана «надкорпусная» База данных лингвоспецифичных единиц русского языка, входом которой являются межъязыковые соответствия, или *моноэквиваленции*, т. е. пары вида: «фрагмент текста, содержащий лингвоспецифичную единицу — функционально-эквивалентный фрагмент текста перевода». Множество тестируемых лингвоспецифичных единиц было сформировано на основе Указателя лексем к книге [Зализняк, Левонтина, Шмелев 2012].

В исследовании применяется *унидирекциональный* метод контрастивного анализа. Он основан на том, что сопоставление двух языков является не целью, а лишь инструментом анализа одного из них: перевод (т. е. решение, реально принятое переводчиком) некоторой лексической единицы русского языка рассматривается как источник сведений о его семантике. Причем в этих целях используется как прямой, так и обратный перевод: соответственно, в случае обратного перевода в качестве свидетельства о семантике анализируемой единицы русского языка рассматриваются условия ее появление в переводе.

Остановимся подробнее на понятии «лингвоспецифичных слов».

Этот термин восходит к работам Анны Вежбицкой, которая говорит о “universal human concepts in culture-specific configurations”, а также о “language-specific meanings” и “language-specific word-meanings”, см. [Wierzbicka 1992, 1996]. В наших работах «лингвоспецифичными» называются слова, заключающие

в себе уникальную концептуальную конфигурацию, т. е. такую, которая «в готовом виде» не представлена ни в какой языковой единице других языков. В этом смысле они непереводимы: при переводе недостающий семантический компонент либо вводится при помощи дополнительных лексических средств, либо просто теряется. При этом, как показал наш опыт семантического анализа, смысловые компоненты, обусловливающие такую труднопереводимость некоторых русских слов, образуют определенную систему, а именно, обычно оказывается, что они отсылают к одной из «ключевых идей», или «сквозных мотивов» русской языковой картины мира, т. е. к некоторым смыслам, которые являются выделенными в рамках семантической системы русского языка[2]. Поэтому в нашей книге [Зализняк, Левонтина, Шмелев 2005] лингвоспецифичные слова определяются двояким образом: с одной стороны, это слова, для которых трудно найти переводной эквивалент; с другой стороны, это слова, которые включают смысловые компоненты, отсылающие к «ключевым идеям».

Однако эти два признака (отсутствие однозначного переводного эквивалента и наличие некоторого семантического компонента, отсылающего к какой-либо «ключевой идее») априори все же задают два разных класса единиц; при этом ни пересечение, ни объединение этих множеств не дает результата, который бы не противоречил интуиции[3]. Поэтому более корректным представляется в качестве определения лингвоспецифичности использовать собственно семантический критерий, а оба указанных обстоятельства считать характерными, но не ингерентными признаками лингвоспецифичности.

Прежде всего, следует напомнить, что лингвоспецифичность — свойство, по своей природе компаративное; при этом, поскольку сравнение со всеми существующими языками (в том числе, в том приближении, которое принято в типологии, когда делается представительная произвольная выборка из языков разных семей и ареалов) не входит в нашу задачу, сравнение производится в рамках актуальной для русского языка оппозиции с основными языками Западной Европы (английский, немецкий, французский, итальянский, испанский.

Как известно, однозначный перевод имеют только термины, и то не всегда. Слова естественного языка в большинстве своем многозначны, и чаще всего в разных значениях и в разных контекстах имеют разные переводные эквиваленты. Кроме того, значение слова естественного языка включает в себя, помимо собственно толкования, также смысловые компоненты, вносимые его функционально-стилистической характеристикой, внутренней формой, эпидигматическими связями, интертекстуальной и культурной нагрузкой и т. д. Очевидно, что

---

[2] А. Вежбицкая называет их «доминантами семантического универсума» русского языка (“semantic themes which shape the semantic universe of the Russian language” [Wierzbicka 1992: 395].

[3] Пересечение, т. е. выполнение обоих требований, оставляет за пределами искомой категории такие слова как *разлука* (безусловно лингвоспецифичные, хотя и имеющие более менее однозначный переводной эквивалент, см. ниже), а также достаточно многочисленную категорию труднопереводимых слов, в значении которых отсылка к какой-либо «ключевой идее» по меньшей мере проблематична (ср. напр. нем. *sich melden*); с другой стороны, объединение, т. е. достаточность какого-то одного из двух признаков, включает туда всю безэквивалентную лексику, что также искажает исходный замысел.

при учете всех этих обстоятельств лингвоспецифичным, т. е. содержащим уникальный набор значений перечисленных параметров, окажется практически любое слово любого языка (вспомним, какого труда стоит А. Вежбицкой выявление «экспонентов» семантических примитивов, и результат все равно является не бесспорным). Тем не менее, категория лингвоспецифичных слов представляется осмысленной[4]: в любом случае, безусловно существует *мера лингвоспецифичности*; именно ее хотелось бы научиться определять.

Для начала выделим следующие четыре основных типа отсутствия межъязыкового семантического изоморфизма (на примере пары «русский — английский»; каждый тип обозначается реализующим его русским словом; в скобках указывается информация об английском переводном эквиваленте):

I) a. подтип РУКА (англ. *hand* vs. *arm*);
   b. подтип ПРАВДА-ИСТИНА (англ. *truth*);
II) тип САМОВАР (переводной эквивалент отсутствует; используется заимствованное слово или описательное определение);
III) тип РАЗЛУКА (имеется один преимущественный вариант перевода, но он неточный);
IV) тип ТОСКА (имеется несколько приблизительно равновероятных вариантов перевода, все неточные).

Все эти типы слов русского языка являются в том или ином смысле лингвоспецифичными. Но это их свойство представляет интерес с разных точек зрения. Первый тип составляет предмет *лексической типологии*, которая устанавливает возможные способы членения предметных или концептуальных областей за счет различных способов колексификации, выбираемых разными языками (см., в частности, [François 2008, Рахилина, Резникова 2013]).

Второй тип — это так называемая *безэквивалентная лексика* (куда относятся, прежде всего, обозначения культурно-специфических реалий, типа *тамада* или *кимоно*), которая много обсуждается в теории перевода. Главная особенность безэквивалентной лексики состоит в том, что она чаще всего подвергается заимствованию — в результате чего, слово, естественно, утрачивает свой безэквивалентный статус в языке-источнике. Однако часто бывает так, что при заимствовании значение слова сдвигается, и межъязыковая эквивалентность снова нарушается (и в этом случае может возникнуть собственно лингвоспецифичность, ср. слова *азарт, кураж, гонор* [Шмелев 2002], а также новые заимствования типа *имидж, пиар*).

Для нас наибольший интерес представляют две последних категории, именно о них и будет идти дальше речь.

---

[4] Отметим, что категория лингвоспецифичных слов имеет прочный аналог в обыденном сознании, ср. популярный Интернет-жанр «10 непереводимых слов»: в этих списках фигурируют, в частности, такие слова как чеш. *lítost*, нем. *Schadenfreude*, порт. *saudade*, русск. *тоска*.

Итак, лингвоспецифичной мы будем называть лексическую единицу, заключающую в себе ***уникальную концептуальную конфигурацию***, а именно такую, что во всех существующих ее переводных эквивалентах не хватает какого-то семантического компонента или неустранимым образом присутствует «лишний» компонент.

Так, например, русское слово *бабушка* в значении термина родства имеет встроенный компонент «взаимных добрых чувств» (см. [Wierzbicka 1992: 351]), который порождает специфическую концептуальную конфигурацию[5]. Данный смысловой компонент отсутствует в англ. *grand-mother*, франц. *grand-mère*, нем. *Großmutter*, при том, что это единственный и конвенциональный переводной эквивалент. В английском языке помимо *grand-mother* есть еще *gran*, *granny*, в немецком есть слово *Oma*, но и то, и другое — это именно уменьшительно-ласкательные обозначения, а русск. *бабушка* не является уменьшительно-ласкательным, это нейтральное слово; по крайней мере, никакого другого слова нет: слово *баба* не имеет нужного значения, кроме как в «детском языке» в сочетании с именем (*Баба Галя*), языке слово *бабка* в стандартном литературном стилистически маркировано как имеющее грубоватый оттенок.

Другой пример: глагол *успеть*, ср.:

(1) Когда это он ***успел*** опять лечь-то! (И. А. Гончаров «Обломов»)
— Quand est-ce qu'il ***a trouvé le temps*** de se recoucher?

Французский эквивалент *trouver le temps* вносит компонент планирования своего времени; между тем, в русском оригинале речь идет не столько о возможной нехватке времени, сколько о наклонности и способности, с одной стороны, и о случайности и удаче — с другой (см. [Зализняк, Левонтина 1996]).

Назовем теперь несколько характерных признаков, указывающих на вероятную лингвоспецифичность слова[6]; некоторые из них могут быть установлены при помощи автоматических процедур, применяемых к параллельным корпусам. А именно, если мы рассматриваем некоторую лексическую единицу (ЛЕ) интересующего нас языка, то ее лингвоспецифичность тем более вероятна:

- чем больше имеющих приблизительно равную частотность «моделей перевода»[7] она имеет;

---

[5] Лингвоспецифичность слова бабушка подтверждается наличием у него знаковой функции в обыденном сознании, ср. его использование в фильме Карена Шахназарова «Американская дочь» в качестве кодового слова, отсылающего к русскому культурному скрипту.

[6] Эти признаки представляют собой развитие принципов квантитативного метода оценки лингвоспецифичности, предложенного Д. В. Сичинавой в докладе на II Международном симпозиуме по лексической типологии (Гранада, 17–19 сентября 2012), см. [Сичинава 2015 (в печати)].

[7] Термины «модель перевода» (слово или словосочетание, используемое в качестве переводного эквивалента рассматриваемой лексической единицы) и «стимул перевода» (слово или словосочетание, в качестве «реакции» на которое появляется интересующая нас языковая единица в обратном переводе) введены в [Loiseau et al. 2013].

- чем более неоднословными они являются;
- чем чаще при переводе происходит замена части речи (напр. существительному соответствует глагол);
- чем больше количество имеющих приблизительно равную частотность «стимулов перевода»[8];
- чем более неоднословными являются эти «стимулы»;
- чем в большем проценте случаев в иностранном оригинале вообще отсутствует какая-либо единица-«стимул», которую можно было бы сопоставить анализируемой русской ЛЕ (ср. ниже примеры на слово *разлука*);
- чем больше расхождений имеется между списками моделей и стимулов перевода данной ЛЕ;
- чем более частотной является исследуемая ЛЕ по сравнению с ее моделями перевода (ср. о сравнительной частотности слов русск. *душа* и англ. *soul* в [Wierbicka 1992]).

Первый из перечисленных принципов требует специального комментария, поскольку он является одной из наиболее популярных мишеней критики идеи непереводимости и стоящей за ней лингвоспецифичности. Так, в статье [Павлова, Безродный 2013: 154] приводится 13 вариантов перевода на немецкий язык русского слова *пошлый*: *kitschig, ordinär, vulgär, gewöhnlich, geschmacklos, niveaulos, primitiv, beschränkt* и др. Заметим, что почти все эти или аналогичные им слова есть и в русском языке (*китч*, *вульгарный*, *безвкусный*, *низкопробный*, *примитивный* и т. д. — однако ни одно из них не содержит ту уникальную концептуальную конфигурацию, которая заключена слове *пошлый* (см. об этом слове подробно [Зализняк, Левонтина, Шмелев 2005: 175–204]). Действительно, установить эти различия позволяет только собственно семантический анализ, однако наличие множества единиц, реально используемых переводчиками и имеющих приблизительно равную частотность — это сигнал о том, что, возможно, данное слово не имеет ни одного точного перевода. Иными словами, наличие большого количества возможных переводных эквивалентов некоторой языковой единицы не опровергает, а скорее подтверждает ее лингвоспецифичность — хотя и не является окончательным аргументом, поскольку, как уже было сказано, наличие множества переводных эквивалентов может быть обусловлено разными причинами.

Проиллюстрируем применение некоторых из перечисленных выше принципов установления меры лингвоспецифичности языковой единицы на двух примерах, иллюстрирующих два разных типа лингвоспецифичных слов.

---

[8] По данным из [Сичинава 2015 (в печати)] в текстах англо-русского параллельного подкорпуса НКРЯ слово *удаль* встретилось 11 раз: два стимула встретились дважды: *prowess, recklessness*, девять — по одному разу: *abandonment, assurance, violence, effrontery, feats, bravado, fearless, confidence*. Для слова *тоска* (одно из самых известных лингвоспецифичных русских слов, см. в частности [Wierzbicka 1992, Шмелев 2002]) в англо-русском подкорпусе НКРЯ было обнаружено 66 различных «стимулов», из которых половина встречается по 1 разу.

собираться

Глагол *собираться* — один из наиболее ярких представителей категории русских лингвоспецифичных слов (см.[Зализняк, Левонтина 1996], [Зализняк 2006: 209–216]). Этот глагол имеет широкий спектр значений; нас будет интересовать только значение намерения (*собираюсь пойти в кино*; *собираюсь в гости, на концерт, к Маше*). Как показано в [Зализняк 2006: 212–213], выражаемое этим глаголом намерение имеет двоякую концептуализацию: оно может быть представлено как состояние и как процесс, и лингвоспецифичным является только второе. Противопоставление это проявляется, в частности, в аспектуальном поведении. Пара *собраться — собираться* (*жениться, покупать дачу, ехать в Америку*) — это обычная видовая пара с перфектным соотношением ‘переход в состояние’ — ‘результирующее состояние’, характерным для ментальных глаголов, ср. *понять — понимать, почувствовать — чувствовать, поверить — верить, показаться — казаться, огорчиться — огорчаться* и т. п. В этом значении глагол *собираться* — ничем не выдающийся, он легко заменяется на *намереваться*, *иметь намерение* и даже иногда *планировать* (ср. *Я собираюсь/намереваюсь/имею намерение/планирую поехать на конференцию по аспектологии в Гётеборге*). Соответственно, столь же безболезненно он переводится при помощи иноязычных эквивалентов: англ. *be going to, plan to, be planning to, intend to*; франц. *s’apprêter à, se préparer à, se proposer, avoir intention, aller, vouloir*; нем. *vorhaben, Absicht haben, wollen* и т. д.

Лингвоспецифичное «квазипроцессное» значение (описанное в [Зализняк, Левонтина 1996]) — это тоже намерение, но представленное как своего рода процесс, который может завершиться или не завершиться результатом — т. е. выполнением задуманного. Классическим образцом носителя этого когнитивного сценария является Илья Ильич Обломов; ср. следующее его детальное описание:

(2) Он, как только проснулся, тотчас же *вознамерился* встать, умыться и, напившись чаю, подумать хорошенько, кое-что сообразить, записать и вообще заняться этим делом как следует. *С полчаса он всё лежал, мучась этим намерением*, но потом рассудил, что *успеет* ещё сделать это и после чаю, а чай можно пить, по обыкновению, в постели, тем более, что ничто не мешает думать и лёжа. Так и сделал. После чаю он уже приподнялся с своего ложа и чуть было не встал; поглядывая на туфли, он даже начал спускать к ним одну ногу с постели, но тотчас же опять подобрал её.

В таком «квазипроцессном» значении глагол *собираться* образует с *собраться* предельную (квазипредельную) пару: можно *собираться и наконец собраться* (ср. *строили, строили и наконец построили*), но можно *собираться, но так и не собраться*. В этом значении обсуждаемый глагол труднопереводим — в том смысле, что он переводится теми же глаголами, но с потерей процессной составляющей и компонента неконтролируемости.

В нашей Базе данных лингвоспецифичных единиц русского языка такое *собираться* представлено в следующих примерах:

(3) — *Давно* ***собирался*** к тебе, — говорил гость, — да ведь ты знаешь, какая у нас дьявольская служба! (И. А. Гончаров. Обломов)
— Depuis longtemps je ***m'apprêtais*** à te rendre visite. Mais tu connais le travail de tous les diables que nous avons.
— Il y a déjà longtemps que je ***me préparais*** à venir te voir, mais tu connais ce damné service que nous avons!

(4) — А я так и не был у него с самых праздников. *Все* ***собирался***. (Л. Н. Толстой. Смерть Ивана Ильича)
— Et moi qui ne suis pas allé chez lui depuis les fêtes.Cependant, ***je m'étais juré de le faire***.

(5) геморрой-с... *всё* гимнастикой ***собираюсь*** лечиться; там, говорят, статские, действительные статские и даже тайные советники охотно через веревочку прыгают-с; (Ф. М. Достоевский. Преступление и наказание)
— des hémorroïdes, vous savez... ***je me propose*** *toujours* de les soigner par la gymnastique; on dit que des conseillers d'Etat sautent volontiers à la corde

Как мы видим, во всех четырех случаях употреблены разные глаголы (в то время как переводные эквиваленты для неспецифичного стативного *собираться* воспроизводятся одни и те же: с точки зрения частотности с большим отрывом лидируют строевые глаголы *aller* и *vouloir*). При этом все четыре французских переводных эквивалента специфического «квазипроцессного» значения не выражают — хотя в остальном перевод достаточно точно передает смысл.

Особенно характерен для рассматриваемого значения употребление в контексте (*собирался, но*) *так и не собрался*. Интересно, что оно довольно часто появляется в переводах на русский язык, в качестве реакции на разнообразные многокомпонентные стимулы, ср. (примеры из НКРЯ):

(6) Throughout that time he had been intending to alter the name over the window, but *had never quite got to the point of doing it*. [George Orwell. Nineteen Eighty-Four (1949)]
— Все эти годы он собирался сменить вывеску, ***но так и не собрался***. [Джордж Оруэлл. 1984 (В. Голышев, 1989)]

(7) Although she had thought in New York, when Eugene first began to make money, that now she would indulge in tailor-made garments and the art of an excellent dressmaker, she *had never done so*. [Theodore Dreiser. The "Genius" (1915)]
— В Нью-Йорке, когда Юджин стал хорошо зарабатывать, она мечтала о том, что будет заказывать наряды у лучших портних, но ***так и не собралась*** этим заняться. [Теодор Драйзер. Гений (М. Волосов, 1930)]

(8) She *never seemed to say* what she had intended to say to the prisoner. [Ursula Le Guin. The Tombs of Atuan (1971)] [омонимия не снята]
— Но кажется, ***так и не собралась*** высказать ему все, что хотела. [Урсула Ле Гуин. Гробницы Атуана (И. Тогоева, 1991)]

(9) Ich bin *noch nicht dazu gekommen*, mir schwarze zu kaufen, oder vielmehr, ich habe es unterlassen. [Thomas Mann. Buddenbrooks (1896–1900)]
— Я еще ***не собрался*** купить черные, или, вернее, решил не покупать. [Томас Манн. Будденброки (Н. Ман, 1953)]

(10) Tu' ich 's jetzt nicht, *so geschäh' es niemals*. [Johann Wolfgang Goethe. Die Leiden des jungen Werther (1774)]
— Если не сейчас, я ***не соберусь никогда***. [Иоганн Вольфганг Гёте. Страдания юного Вертера (Н. Касаткина, 1954)]

Как можно видеть, из всех этих примеров лишь в двух случаях «стимул перевода» содержит идею 'не собрался', и оба они неоднословны: англ. *had never quite got to the point of doing it* в (6) и нем. *noch nicht dazu gekommen* в (8). В остальных случаях русское *не собрался* возникает в качестве эквивалента для единиц со значением 'не сделал', как бы сглаживая более «жесткие» очертания английского смысла за счет идеи неполной контролируемости. Пример (9) обнаруживает компонент неконтролируемости русского *не собрался* наоборот, за счет того, что в качестве стимула перевода использована конструкция без личного субъекта.

разлука

Слово *разлука* (проанализированное в [Зализняк, Левонтина 1999], [Levontina, Zalizniak 2001]) имеет в значение 'временное пребывание в состоянии пространственной разделенности с кем-то, сопровождающееся специфическим эмоциональным состоянием'. Это слово имеет конвенциональные переводные эквиваленты: англ. *separation*; франц. *séparation*, нем. *Trennung*, в которых, однако, отсутствует компонент эмоционального состояния, встроенный в русское слово *разлука* и составляющий его специфику.

В русско-французском параллельном подкорпусе НКРЯ слово *разлука* встречается 30 раз. Во французском переводе им соответствует: *séparation* (19), *être séparé* (2), *se séparer* (2), *absence* (2); описательный способ передачи смысла: *sans toi* (*la vie est trop affreuse*) и др. (5). Во французско-русском подкорпусе слово *разлука* встречается в пяти примерах, все пять имеют разные «стимулы»: *se quitter, être séparé, éloignement, privation, absence.*

Еще более показательны данные англо-русского параллельного подкорпуса (в несколько раз большего по объему). В переводах с английского на русский в качестве «стимула», помимо слова *separation* могут выступать: *absence (from smbody), severance, leaving, to be absent (from smbody), to be separated (from smbody), to be parted (from somebody), being away (from smbody), being away long*

enough/too long, grief of losing, not seeing him soon again, good-bye (That good-bye had lasted until now) и другие относительно свободные словосочетания. В оригинале может вообще не быть никакого одного слова, которое служит стимулом; слово *разлука* в русском переводе возникает в как аккумулятор смыслов ‘отсутствие контакта’ и ‘эмоциональное состояние, включающее страдание от отсутствия контакта и желание его возобновления’, присутствующих в предложении в разных местах и выраженных различными способами. Ср. (фрагменты английской фразы, послужившие «стимулом» для появления слова *разлука* в переводе, выделены курсивом):

(11) Scarlett, seeing him for *the first time in more than two years*, was frightened by the *violence of her feelings*. [Margaret Mitchell. Gone with the Wind, Part 1 (1936)]
— Буря чувств, которую эта встреча, первая после двух лет ***разлуки***, пробудила в душе Скарлетт, потрясла и испугала ее самое. [Маргарет Митчелл. Унесённые ветром, ч. 1 (Т. Озерская, 1982)]

(12) (12) But there were some nights when even brandy would not still the ache in her heart, the ache that was even stronger than fear of losing the mills, the *ache to see* Tara *again*. [Margaret Mitchell. Gone with the Wind, Part 2 (1936)]
— Но бывали ночи, когда даже с помощью коньяка Скарлетт не удавалось утишить боль в сердце, боль, куда более сильную, чем страх потерять лесопилки, — неутолимую боль ***разлуки*** с Тарой. [Маргарет Митчелл. Унесённые ветром, ч. 2 (Т. Кудрявцева, 1982)]

Ср. также следующий пример перевода с русского:

(13) Сколько месяцев не слышал паровозного крика, и как моряка волнует бирюзовая синь бескрайнего моря каждый раз после долгой ***разлуки***, так и сейчас кочегара и монтера звала к себе родная стихия. [Н. А. Островский. Как закалялась сталь (ч. 2) (1930–1934)]
— It was months since he had heard an engine whistle, and the one-time stoker and electrician yearned as much for the familiar surroundings as the sailor *yearns* for the boundless sea expanse after *a prolonged stay on shore*. [Nikolai Ostrovsky. How the Steel was Tempered (pt 2) (R. Prokofieva, 1952)]

Здесь «модель перевода» слова *разлука* складывается из двух фрагментов: со значением ‘долгое нахождение в другом месте’ (*a prolonged stay on shore*) и со значением определенного эмоционального состояния (*yearns* ‘страстно стремится’).

Таким образом, применение предлагаемых принципов оценки степени лингвоспецифичности лексической единицы методом унидирекционального контрастивного анализа подтверждает полученные ранее результаты собственно семантического анализа и в дальнейшем — при условии значительного увеличения объема параллельных корпусов — может использоваться как метод выявления гипотетически лингвоспецифичных языковых единиц.

Автор пользуется случаем выразить благодарность анонимным рецензентам за ценные замечания, которые по возможности были учтены в окончательной версии статьи.

Литература

1. *Бунтман Н. В., Зализняк Анна А., Зацман И. М., Кружков М. Г., Лощилова Е. Ю., Сичинава Д. В.* (2014), Информационные технологии корпусных исследований: принципы построения кросс-лингвистических баз данных // Информатика и ее применения. Т. 8, вып. 2. С. 98–110.
2. *Зализняк, Анна А.* (2006), Многозначность в языке и способы ее представления. М.: «Языки славянских культур».
3. *Зализняк Анна А.* (2013), Русская семантика в типологической перспективе. К вопросу о термине «языковая картина мира» // Russian Linguistics, Vol. 37, Issue 1, P. 5–20.
4. *Зализняк Анна А.* (2014), Языковая картина мира: между Сциллой и Харибдой // V Международный конгресс исследователей русского языка. Москва, МГУ, 18–21 марта 2014. С. 108–109.
5. *Зализняк Анна А., Левонтина И. Б.* (1996), 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 Univ. Press, 1992) // Russian Linguistics, vol. 20, pp. 237–264.
6. *Зализняк Анна А., Левонтина И. Б.* (1999), С любимыми не расставайтесь // Логический анализ языка: Образ человека в культуре и языке. М.
7. *Зализняк Анна А., Левонтина И. Б., Шмелев А. Д.* (2005) Ключевые идеи русской языковой картины мира. Москва: Языки славянской культуры. М.
8. *Зализняк Анна А., Левонтина И. Б., Шмелев А. Д.* (2012) Константы и переменные русской языковой картины мира. Москва: Языки славянской культуры. М.
9. *Павлова А. В., Безродный М. В.* (2013), Хитрушки и единорог: из истории лингвонарциссизма // От лингвистики к мифу: Лингвистическая культурология в поисках «этнической ментальности». СПб., с. 138–159.
10. *Рахилина Е. В., Резникова Т. И.* (2013), Фреймовый подход к лексической типологии // Вопросы языкознания, № 2, с. 3–31.
11. *Сичинава Д. В.* (2015, в печати) Параллельные тексты в составе Национального корпуса русского языка: новые направления развития и результаты // Труды Института русского языка РАН. М.
12. *Шмелев А. Д.* (2002), Русская языковая модель мира. Опыт словаря. М.: «Языки славянской культуры».
13. *Шмелев А. Д.* (2014), Язык и культура: есть ли точки соприкосновения? // Труды Института русского языка им. В. В. Виноградова. Т. I. М. С 36–116.
14. *François A.* (2008) Semantic maps and the typology of colexification. // From Polysemy to semantic change. Towards a Typology of Lexical Semantic

Associations. Ed. by Martine Vanhove. [Studies in Language Companion series, 106] Amsterdam: John Benjamins Publishing Company. P. 163–216.
15. *Kruzhkov M., Buntman N. V., Loshchilova E. J., Sitchinava D. V., Zalizniak Anna A., Zatsman I. M.* (2014), The database of Russian verbal forms and their French translation equivalents // Компьютерная лингвистика и интеллектуальные технологии. Труды Международной конференции Диалог'2014. С. 275–287.
16. *Levontina I. B., Zalizniak Anna A.* (2001), Human emotions viewed through the Russian language. // Emotions in Crosslinguistic perspective. Ed. by J. Harkins and A. Wierzbicka Berlin — N.Y.: Mouton de Gruyter, 2001, pp. 291–336.
17. *Loiseau S., Sitchinava D. V., Zalizniak Anna A., Zatsman I. M.* (2013), Information technologies for creating the database of equivalent verbal forms in the Russian-French multivariant parallel corpus // Информатика и ее применения, 2013. Том 7, вып. 2. С. 100–109.
18. *Wierzbicka A.* (1992), Semantics, Culture, and Cognition. Universal Human Concepts in Culture-Specific Configurations. N. Y.; Oxford: Oxford Univ. Press.
19. *Wierzbicka A.* (1996), Semantics: Primes and Universals. Oxford: Oxford Univ. Press, 1996.

References

1. *Buntman N. V., Zalizn jak Anna A., Zatsman I. M., Kruzhkov M. G., Loshchilova E. Ju., Sitchinava D. V.* (2014), Informacionnye tekhnologii korpusnykh issledovanij: printsipy postroenija kross-lingvisticheskikh baz dannykh // Informatika i ee primenenija. T. 8, vyp. 2. S. 98–110.
2. *François A.* (2008), Semantic maps and the typology of colexification. // From Polysemy to semantic change. Towards a Typology of Lexical Semantic Associations. Ed. by Martine Vanhove. [Studies in Language Companion series, 106] Amsterdam: John Benjamins Publishing Company. P. 163–216.
3. *Kruzhkov M., Buntman N. V., Loshchilova E. J. Sitchinava D. V., Zalizniak Anna A., Zatsman I. M.* (2014), The database of Russian verbal forms and their French translation equivalents // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2014". S. 275–287.
4. *Levontina I. B., Zalizniak Anna A.* (2001), Human emotions viewed through the Russian language. // Emotions in Crosslinguistic perspective. Ed. by J. Harkins and A. Wierzbicka Berlin — N.Y.: Mouton de Gruyter, 2001, pp. 291–336.
5. *Loiseau S., Sitchinava D. V., Zalizniak Anna A., Zatsman I. M.* (2013), Information technologies for creating the database of equivalent verbal forms in the Russian-French multivariant parallel corpus // Onformatika I ee primrnenija, 2013. T. 7, vyp. 2. S. 100–109.
6. *Pavlova A. V., Bezrodnyj M. V.* (2013), Hitrushki i edinorog: iz istorii lingvonartsissizma // Ot lingvistiki k mifu: Lingvisticheskaja kul'turologija v poiskakh «etnicheskoj mental'nosti». SPb., s. 138–159.
7. *Rahilina E. V., Reznikova T. I.* (2013), Frejmovyj podkhod k leksicheskoj tipologii // Voprosy jazykoznanija, №2, s. 3–31.

8. *Sichinava D. V.* (2015, in print), Parallel'nye teksty v sostave Natsional'nogo korpusa russkogo jazyka: novye napravlenija razvitija i rezul'taty //Trudy Instituta russkogo jazyka RAN. M.
9. *Shmelev A. D.* (2002), Russkaja jazykovaja model' mira. Opyt slovarja. M.: «Jazyki slavjanskoj kul'tury».
10. *Shmelev A. D.* (2014), Jazyk i kul'tura: est' li tochki soprikosnovenija? // Trudy Instituta russkogo jazyka im. V. V. Vinogradova. T. I. M. S 36–116.
11. *Wierzbicka A.* (1992), Semantics, Culture, and Cognition. Universal Human Concepts in Culture-Specific Configurations. N. Y.; Oxford: Oxford Univ. Press.
12. *Wierzbicka A.* (1996), Semantics: Primes and Universals. Oxford: Oxford Univ. Press, 1996.
13. *Zaliznjak, Anna A.* (2006), Mnogoznachnost' v jazyke i sposoby ee predstavlenija. M.: «Jazyki slavjanskih kul'tur».
14. *Zaliznjak Anna A.* (2013), Russkaja semantika v tipologicheskoj perspektive. K voprosu o termine «jazykovaja kartina mira» // Russian Linguistics, Vol. 37, Issue 1, P. 5–20.
15. *Zaliznjak Anna A.* (2014), Jazykovaja kartina mira: mezhdu Scilloj i Kharibdoj // V Mezhdunarodnyj kongress issledovatelej russkogo jazyka. Moskva, MGU, 18–21 March 2014. S. 108–109.
16. *Zaliznjak Anna A., Levontina I. B.* (1996), Otrazhenie natsional'nogo kharaktera v leksike russkogo jazyka (razmyshlenija po povodu knigi: Anna Wierzbicka. Semantics, Culture, and Cognition. Universal Human Concepts in Culture-Specific Configurations. N.Y., Oxford, Oxford Univ. Press, 1992) // Russian Linguistics, vol. 20, pp. 237–264.
17. *Zaliznjak Anna A., Levontina I. B.* (1999), S ljubimymi ne rasstavajtes' // Logicheskij analiz jazyka: Obraz cheloveka v kul'ture i jazyke. M.
18. *Zaliznjak Anna A., Levontina I. B., Shmelev A. D.* (2005), Kljuchevye idei russkoj jazykovoj kartiny mira. Moskva: Jazyki slavjanskoj kul'tury. M.
19. *Zaliznjak Anna A., Levontina I. B., Shmelev A. D.* (2012), Konstanty i peremennye russkoj jazykovoj kartiny mira. Moskva: Jazyki slavjanskoj kul'tury. M.

ЯЗЫКИ V2: ВТОРАЯ ПОЗИЦИЯ И ПЕРЕДВИЖЕНИЕ ГЛАГОЛА

Циммерлинг А. В. (fagraey64@hotmail.com)
ИСЛИ МГГУ им. М. А. Шолохова;
Институт языкознания РАН, Москва, Россия

Лютикова Е. А. (lyutikova2008@gmail.com)
МГУ им. М. В. Ломоносова;
ИСЛИ МГГУ им. М. А. Шолохова, Москва, Россия

В статье обсуждаются конституирующие свойства V2-языков с точки зрения параметрической типологии. V2-языки представляют собой группу языков со сходной архитектурой клаузы и, в частности, единой второй позицией финитного глагола. В статье обосновывается утверждение, что «классический» формальный анализ V2, связанный с передвижением глагола и особым набором признаков у вершины, вызывающим такое передвижение, эмпирически адекватен, а современные попытки расширить круг V2-языков за счет ослабления диагностических требований контрпродуктивны. Языки с «частичным» или «остаточным» V2, определяемые по позиции глагола лишь в части независимых утвердительных предложений, не образуют естественного класса в отношении прочих параметров структуры клаузы; в то же время передвижение глагола само по себе не является достаточным условием для возникновения V2.

Ключевые слова: языки V2, передвижение, синтаксис, параметрическая типология, коммуникативная структура, парсинг

APPROACHING V2: VERB SECOND AND VERB MOVEMENT[1]

Anton Zimmerling (fagraey64@hotmail.com)
Institute for Modern Linguistic Research SMSUH;
Institute of Linguistics, Russian Academy of Sciences,
Moscow, Russia

Ekaterina Lyutikova (lyutikova2008@gmail.com)
Lomonosov Moscow State University;
Institute for Modern Linguistic Research SMSUH,
Moscow, Russia

[1] This research has been supported by the Russian Science Foundation, project RFH 14-18-03270 'Word order typology, communicative-syntactic interface and information structure in world's languages'. The authors are grateful to the anonymous reviewer for the valuable critical comments.

The paper discusses constituting properties of V2 languages in a perspective of parametric typology. V2 languages are a small group of syntactically uniform languages sharing a number of parameters constraining the clausal architecture and the finite verb placement. We argue that whereas the generative procedure of deriving V2 by verb movement and feature composition of the target head is correct and has empirical validation, the broader definitions of V2 phenomena found in the contemporary work on the subject that loosen the diagnostic criteria on the single preverbal constituent are counterproductive. So called 'partial' or 'residual' V2 languages, where verb movement to the left-peripheral position is allegedly characteristic for a part of root declaratives, do not exist; at the same time, the verb movement by itself is not sufficient to produce the classic V2 profile.

Key words: verb-second languages, movement, syntax, parametric typology, information structure, parsing

1. V2 languages: data, framework and NLP parsing

The term 'Verb-Second Language' (V2 languages) refers to a relatively small class of world's languages where finite verbs have a fixed position in some types of declarative clauses. Following (Zimmerling 2002), we distinguish between 'Strict V2 languages', e.g. German, Dutch, Afrikaans, Danish, Swedish, Norwegian, where the verb invariably takes the clause-second position, and 'V1/V2 languages', e.g. Old and Modern Icelandic, Faroese, Middle Norwegian, Old High German, Yiddish, where both second and first positions (but not the 3rd, the 4th, etc.) are licensed. Almost all strict V2 and V1/V2 languages belong to the Germanic group of Indo-European languages, the notable exceptions being Kashmiri and Raeto-Romance, cf. Kaiser (2002). V2 is a grammaticalized formal constraint which seems to be independent from information structure (IS) and prosodic issues.

V2 languages share many parameter settings in word order, which can be implemented in rule-based NLP parsing, cf. Vikner (1995), Bhatt (1999) and Wolfe (2015). Generative accounts of V2, e.g. den Besten (1983), Holmberg & Platzack (1995) and Holmberg (2015) work out the insight that verb-second phenomenon is not a primitive feature of a language but a superficial generalization on the clause structure that is triggered by more general mechanisms, notably—verb movement to a dedicated position in the left periphery and movement of phrasal categories to a specifier position preceding the target position of the moved verb, the two instances of movement being in principle independent from each other. As a consequence, broader definitions of V2 have been proposed that generalize about various instances of clause-internal verb movement. In this paper we argue that although V2 indeed involves verb movement and thus linking V2 to movement of a verb to a dedicated position is correct, V2 is more complex than a side effect of verb movement.

2. Definitional properties of V2

V2 languages are for syntactic systems fitting two requirements: a) finite verb forms (V_{fin}) take the fixed (second) position in some type of declaratives, usually—in root clauses, cf. Den Besten (1983), but sometimes also in embedded declaratives and in some interrogative clauses, cf. Bhatt (1999); b) the clause-initial position in the diagnostic type of clauses is filled by exactly one constituent—the so called bottleneck condition, in terms of Holmberg (2015). On economic reasons a) and b) can be restated as one general condition: if the second position of a finite verb is generalized in some of type of clauses, there can be exactly one preverbal (i.e. first) constituent in this type of clauses. In Diderichsen (1976), requirements a) and b) are viewed as two sides of one basic condition, a similar analysis is proposed in Zimmerling (2002: 221). However, the current practice is keeping a) and b) apart, since verb movement and accompanying movement of phrasal categories to the preverbal position have different motivation (Roberts 2012; Holmberg 2015). In addition, the exact number of clause-internal preverbal constituents is not always clear, since tentative V2 languages can have topicalization constructions, cf. section **2.3** below. Moreover, some languages which ban V>3 in a diagnostic type of declarative clauses at the same time license V1 orders in the same type of clauses, whereby the preverbal position is not filled by any overt sentence material, cf. section **2.5** below. Bech & Salvesen (2014) and Wolfe (2015) claim that the verb can reach its dedicated position in the clausal left periphery even if the bottleneck condition is not satisfied in any type of clauses: such word order systems are called 'residual V2 languages', cf. the discussion in section **3**. In what follows we show that topicalization constructions, V1 orders and multiple XP-fronting can be explained without abandoning the bottleneck condition.

2.1. The bottleneck condition

A descriptive schema of V2 needs three symbols—a symbol of clausal (left) border (#), a symbol of the preverbal constituent (XP) and a symbol for the finite verb (V_{fin}), as represented by G(eneralization) 1.

(G1) #XP—V_{fin}, *#XP—Y—V_{fin}

The bottleneck condition is crucial for V2 diagnostics. It predicts two features of V2 syntax: a) that a combination of two (or more) phrasal categories X, Y preceding the finite verb in a V2 language should be ungrammatical; b) that the XP-position in the diagnostic type of V2 declaratives is not reserved for any particular syntactic category (e.g. noun phrase) and does not express any particular grammatical relation (e.g. subject), hence {-EPP} in Roberts' (2012) terminology. In other words, the XP-position in a V2 language can be filled by any element in an OR-expression $\{Cat_1 \vee Cat_2 \vee \ldots Cat_n\}$, but simultaneous spell-out of two or more hierarchically independent categories in XP is blocked. Parsing

of well-formed V2 structures is licensed by a combination of an OR-expression filter, which lists sentence categories that fill XP in language L, and an &-expression filter which determines which types of expressions count as single constituents when filling XP in language L.

2.2. XP-movement and multiple *wh*-fronting

All V2 languages are sensitive to constituency rules. The regular constraints on the phrasal movement are further adjusted by additional requirements specific for the 1st position promotion. Thus, Danish strongly favours the PP split in clause-first XP promotion, being more tolerate in other instances of phrasal movement (see, e.g., Zimmerling 2002: 234–6):

(1) a. [$_{XP}$ Min mor] har han ikke nok [$_{DP}$ stor tillid [$_{PP}$ til ~~min mor~~]].
my mother have.PRS he NEG enough big confidence to
'He surely does not have a big confidence to my mother.'
b. *[$_{XP}$ Til min mor] har han ikke nok [$_{DP}$ stor tillid ~~til min mor~~].

Another well-known obstacle to dropping special rules for XP-movement is that XP in some V2 languages hosts not only whole constituents (maximal projections), but also parts of them. Old Icelandic examples (2a-b) show extraction of a head element from a DP and left branch extraction (LBE).

(2) a. [$_{XP}$ ∨tfall] var [$_{DP}$ ~~útfall~~ sjávarinnar].
flood.NOM.SG. be.3SG.PRT sea.GEN.SG.DEF
'The was a flood of tide'
b. Þeirrar skal=tu [$_{DP}$ ~~þeirrar~~ konu] biðja.
this.GEN.SG.F shall.2SG=CL.2SG. woman woo
'You shall woo that woman'

Finally, in some cases it is impossible to determine whether a sequence Cat_1 & Cat_2 forms a single constituent or not without checking its capacity to fill XP. So Norwegian, which is considered a strict V2 language, occasionally licenses sequences of several adverbials in XP, cf. (3).

(3) [$_{XP}$ [$_{AdvP}$ I byen] [$_{AdvP}$ i dag]] trefte jeg Marit.
in town today met I Marit
'Today, I met Marit in the town'.

Both adverbials in (3) have the same IS status (correspond to a Theme), and there are no grounds to believe that any of them is extraclausal. Therefore, one must assume that at least those native speakers who accept (3) generate/parse a single adverbial phrase there. Multiple XP-movement is characteristic for a minority of V2 languages including Modern Icelandic, Faroese and Old Swedish, cf. the

discussion in Zimmerling (2002: 291; 501). A similar or identical parameter licenses optional multiple *wh*-fronting in Kashmiri interrogatives, cf. (4).

(4) a. [$_{WhP}$ [$_{WhP}$ Kus] [$_{WhP}$ kemyis] [$_{WhP}$ kyaa]] dii?
who whom what give-FUT
'Who will give what to whom?'
b. [$_{WhP}$ [$_{WhP}$ Kus] [$_{WhP}$ kyaa]] dii [$_{WhP}$ kemyis]?
c. [$_{WhP}$ [$_{WhP}$ Kus]] dii [$_{WhP}$ kemyis] [$_{WhP}$ kyaa] ?

Other V2 languages, e.g. German, ban structures like (3) and (4a), so multiple-XP-movement and multiple *wh*-fronting parameters may take different values in V2 languages. Multiple XP-movement/XP-fronting is compatible with V2, insofar the possibility of making a single constituent out of hierarchically independent phrases is restricted to special types of phrases and to contexts where all components of an ensemble have the same IS value—that of a Theme/Topic in (3) or a Focus/*wh*-word in (4a).

2.3. Topicalization

A group of V2 languages licenses constructions with a left-dislocated topical element coindexed with a main clause element; this is illustrated in (5) from Kashmiri, where the dislocated DP is coindexed with the resumptive pronoun.

(5) [$_{DP}$ Su LaRk]$_i$, Rameshan vuch temis$_i$ tsuur karaan.
that boy.NOM Ramesh.ERG saw he.DAT theft do.N.PERF
'As for that boy, it is Ramesh who saw him stealing'.

Some V2 languages like Swedish also license structures with a dislocated VP-fragment, like that in (6a).

(6) a. [$_{VP'}$ Läser$_j$ boken]$_i$ det$_i$ *gör*$_j$ han nu.
read.3SG.PRS book.SG.DEF it do.3SG.PRS he.NOM now
'He is reading the book, that is what he is doing now'.
b. *[$_{VP'}$ Läser$_j$ boken]$_i$ *gör*$_j$ han nu.
read.3SG.PRES. book.SG.DEF do.3SG.PRS he.NOM now

Note that (6b) with a topicalized verb phrase in XP and a resumptive verb in V2 is ungrammatical in Swedish. Therefore, it is clear that the initial phrase in (6a) is extraclausal (left-dislocated). Structures like (5) and (6a) are compatible with V2, since formal criteria are met confirming that dislocated topics are extraclausal.

2.4. Empirical motivation for verb movement in V2 languages

Framework-internal minimalist accounts of V2 elaborate on the idea that the bottleneck condition is impossible without verb movement to a position in the left periphery. In early versions of the Chomskyan framework this domain has been identified as C, since it was believed that Comp and V_{fin} always have a complementary distribution in V2 languages and compete for one and the same slot, cf. den Besten (1983), Holmberg & Platzack (1995). This seemingly aprioristic claim is based on two empirical generalizations:

(G2) In V2 languages finite and non-finite verbs take different positions.

(G3) In allegedly prototypical V2 languages (Modern German, Dutch, Danish, Swedish, Norwegian) there is root-subordinate clause asymmetry: in the presence of an overt complementizer, finite verbs do not take V2 and are either placed clause-finally—the West-Germanic option, German (7a–b), or one step further to the right, after negation/negative phrases/sentential adverbs—the Mainland Scandinavian option, Danish (8a–b).

(7) a. Der Hans hat dem Peter keine Instruktionen gegeben.
'Hans has not given instructions to Peter.'
b. Ich glaube, [$_{CP}$ **daß** der Hans dem Peter keine Instruktionen gegeben hat].
'I believe that Hans has not given instructions to Peter.'

(8) a. Jens har *ikke* givet instruktioner til Peter.
'Jens has not given instructions to Peter.'
b. Jeg tror, [$_{CP}$ **at** Jens *ikke* har givet instruktioner til Peter].
'I believe that Jens has not given instructions to Peter.'

The generalization G2 predicts that verb movement correlates with finiteness feature and verb morphology (TAM markers and inflectional properties). There are no {-EPP} languages such that their clause-second position is occupied by a base-generated finiteness marker, while their clause-first position is not reserved for a specific syntactic category. Thus, many Mande languages have basic word order S AUX O V and overtly resemble to V2 systems. However, this similarity is superficial, since the clause-initial position is invariably reserved for the grammatical subject.

The generalization G3 is more of a technical issue. It states which type of declarative clauses is diagnostic for a V2 language. In the form given above, (iii) is falsifiable, since there are languages where V2 order comes up not only in root declaratives, but also in some subordinate clauses, e.g. Kashmiri (Bhatt 1999, see also example (9)), Icelandic (Zimmerling 2002: 303) and Afrikaans. Other V2 languages, including Danish, Swedish and Norwegian, have numerous deviations from G3 too, therefore some linguists prefer to speak not of the root *vs*. subordinate clause asymmetry regarding V2, but of 'subordinate clauses with a subordinate clause word order', where the verb does not move, *vs*. 'subordinate clauses with a main clause word order', cf. Vikner (1995).

(9)

	XP	V_{fin}	S	O	V_{inf}			
a.	**raath**	dyut	laRkan	tswaTh	daar-yith.			
	yesterday	gave	boy.ERG	waste.NOM	throw-out			
	'Yesterday the boy threw out waste'.							
	XP	V_{fin}	Comp	XP	V_{fin}	S	O	V_{inf}
b.	tem	dop	[$_{CP}$ **ki**	**raath**	duyt	laRkan	tswaTh	daar-yith].
	'He said that yesterday the boy threw out waste.'							

2.5. V1 and V>2 orders

Different positions of finite verbs in other types of clauses, i.e. V1 orders in interrogatives and conditionals without an overt complementizer, are often considered as an additional proof that finite verbs move in V2 languages. However, V1 orders in interrogatives, imperatives and conditionals lacking an overt complementizer, as well as subject-verb inversion and adjacent verb-subject orders are not diagnostic for V2 languages, contrary to the claims made in Salvesen & Bech (2014). As stated by Kaiser (2002), Kaiser & Zimmermann (2015), these features are widely attested in non-V2 languages that do not fit the 'bottleneck' condition in declaratives, e.g. in Spanish, Italian, French, Middle Romance languages, Basque, Estonian etc. They have different triggers: thus, VS orders in Spanish or Italian are not bound to the presence of a preverbal non-subject constituent, there are varieties of Germanic V2 languages that lack V2 orders in *wh*-questions, etc. Verb-subject adjacency of postverbal subject DPs is not diagnostic either: as noted in Bhatt (1999) and Zimmerling (2002: 490; 2013: 188–195), many V2 languages (German, Dutch, Kashmiri) with scrambling in the middle field (i.e. between V_{fin} and V_{inf}) lack fixed slots for a postverbal subject DP.

2.5.1. V1 orders

V1 orders in *yes-no* questions, imperatives and marginally acceptable V1 declaratives in strict V2 languages are usually explained in generative literature by postulating invisible operators or silent topic elements in XP, cf. Platzack (2008). These are framework-internal explanations characteristic of theories that crucially rely on the assumption that V2 languages always have an overt or silent syntactic category in front of the moved verb. For parametric typology such stipulations are redundant, if one explicitly specifies that in each V2 language V2 orders are restricted to some diagnostic group of clauses, and that in certain clauses the verb moves higher than (the target position of) V2. The functional motivation for this proposal is that V1 clauses have a different illocutionary force than V2 declaratives and it is preferable not to masque this fact by claiming that overt V1 and overt V2 have the same underlying structure. As for V1 declaratives in V1/V2 languages, we raise a stronger claim:

(G4) V1 declaratives in V1/V2 languages are IS-marked and formally derived variants of V2 declaratives.

An analysis of Old Icelandic, Middle Norwegian, Modern Icelandic put forward in Zimmerling (2002: 363–366) shows that V1 declaratives in such V1/V2 systems are found in a wide variety of different contexts, and the tag 'narrative inversion' is just a descriptive convention. V1 declaratives with verb fronting are also found in Russian, cf. Yanko (2001), Zimmerling (2013: 280–283) or Ossetic, cf. Lyutikova & Tatevosov (2009), and the analysis of scrambling patterns in these languages can be easily extended to V1/V2 languages. Indeed, there is no evidence that a ban on V>2 declaratives has any impact on IS-motivated derivation of V1 orders.

2.5.2. V > 2 orders

The specific type of V>2 constructions emerges when the target position of verb movement can be reached by some other sentence category in root clauses. This is a rare option, but it is attested as well. Thus, in Swedish, the modal adverb *kanske* 'maybe' takes the same slot as the tensed verb and competes with it for C/V2, cf. Platzack (2008); very similar Danish (Diderichsen 1976) and Norwegian (Faarlund et al. 1997) word order systems lack this option.

(10) a. Nu **kanske** Johan inte vill komma.
now MAYBE John not FUT come.INF
'John probably won't come now'.
b. Johan **kanske** inte vill komma.

A close parallel to this pattern is found in some Clitic-Second languages like Serbo-Croatian, where V2 orders come up in derived structures with a so called Barrier constituent. With the default word order XP-CL, the clausal-second position is filled by clustering clitics and is of course not available for the verb, cf. (11a). But if the initial topical constituent has Barrier properties, the clitics normally do not attach to it, and the vacant target position is filled by the verb in clauses like (11b), which gives rise to Verb-Second and Clitic-Third orders; see Zimmerling & Kosta (2013: 197–199) and Zimmerling (2013: 445–464) for discussion and further examples.

(11) a. [$_{pp}$ Poslije toga] *=su* dobili pozive u reprezentaciju.
after that CL.AUX3.PL. get.3.PL.PERF calls to national team
'After that, they have been summoned to the national team.'
b. **[^{BARRIER} [$_{pp}$Poslije svega toga]]** bilo *=mi* *=je*
After all that AUX.3SG.N.PERF CL.1SG.DAT CL.AUX.3SG.
potrebno samo ležati na pijesku.
necessary.ADJ.SG.N. only lie.INF on sand
'After all that, everything I needed was to lie on sand.'

All these options are language-specific and subject to microvariation in genetically and areally related V2 idioms. There is, however, one general conclusion we would like to draw. It is not verb movement itself, but the requirement that target position 2P attracting verbs AND/OR some other sentence category must be filled in a diagnostic group of clauses that is crucial for V2 syntax.

3. V2 and cartography

The classic account of V2 in both formal and descriptive grammars captures three basic facts about V2 languages: a) verb movement to a dedicated position is obligatory in the diagnostic group of clauses, b) all categories that can fill XP lie clause-internally, c) head movement to 2P and phrasal movement to SpecTP in V2 and V1/V2 languages have grammar-internal motivation and do not depend on IS/prosody, while marked constructions with V>2 orders have IS-triggers. The revival of the researchers' interest to V2 is due to the cartography hypothesis which suggests a universal template of multiple functional projections arranged in a fixed order common for all languages. According to Rizzi 1997, the left-periphery of the clause has a finer structure like *Force* > *Topic*$_1$ > *Interrogative* > *Topic*$_2$ > *Focus* … *Topic*$_n$ > *Finiteness* > *TP*. The obvious question is, therefore, which one of the multiple projections of a finer-structured left periphery succeeds the non-split single C of the previous analyses in being the locus of V2 phenomena.

It seems that Fin is generally acknowledged as the successor of C; so, Holmberg (2015), Bech & Salvesen (2014), Wolfe (2015) argue that Fin attracts tensed verbs in V2 languages and presumably in a broader class of languages, the so called 'residual' V2 languages, where verb movement to FinP is not generalized in any group of declaratives. As FinP is dominated by other functional projections which can in principle host multiple XPs, the crucial question is whether cartographic theories retain the restrictive 'bottleneck' condition or give it up.

If references to cartography, as in Bhatt (1999: 112), are made just to specify that elements filling XP in an OR-expression {Cat$_1$ ∨ Cat$_2$ ∨ … Cat$_n$} take different slots in the left periphery, some of them being topical, some of them being focal, little if anything changes, except for the claim that XP is a descriptive tag, while exact definitions of its syntactic position come from cartography. The same holds for hybrid accounts of XP-movement, which are based on the idea that only a part of phrasal categories reach the left periphery by movement, while other categories are base-generated there. For instance, Mathieu (2006) argues that only non-subject DPs are moved to the preverbal position in Old French—the language he describes as V2, while subject DPs do not move out of TP. Beninca & Poletto (2004) make a general claim that only focus elements move, while topic elements are base-generated in the left periphery.

If, however, cartographic theories include a claim that both V2 languages and non-V2 languages where the verb moves to the left periphery, but two or more clause-internal categories can precede it, have the same syntactic build-up, multiple issues arise. Sentences with a topical constituent in front of XP are also known in V2 languages, but there, as shown above in **2.3** and **2.5.2**, the topical constituent is extra-clausal, so examples like (5) and (6a), strictly speaking, show not V>2 orders, but V2 orders with a preceding dislocated phrase. For languages like Old English and Old French—which, according to Bech & Salvesen (2014), both have V2—there is no independent verification that any of the preverbal constituents is extraclausal, since the 'bottleneck' condition on a single preverbal phrase is violated in all clause types, cf. <S Adv$_1$ Adv$_2$ O V> order in (12) and <S O$_1$ O$_2$ V O$_3$ > order in (13).

(12) [$_{DP}$ Goliath] [$_{AdvP}$ par quarante jurs] [$_{AdvP}$ le matin é le vesper] [$_{PP}$ a l'ost de Israel] vint
(Bible 1170)
'For forty days, Goliath came to Israel's army in the morning and in the evening'. — Old French.

(13) 7 [hy] [him] [$_{PP}$ æfter ðæm grimme] forguldon
And they them after that cruelty repaid-3PL.PRT
[$_{DP}$ þone wigcræft [$_{CP}$ þe [hy] [$_{PP}$ æt him]] leornodon]. (Or. 22)
that art.of.war that they at him learn-3PL.PRT
'And after that, they bitterly repaid him for the art of war that they learned from him'. — Old English.

To sum up, this updated approach fails to provide effective and reliable criteria for identifying V2. The main problem is that if a language has scrambling of preverbal elements and can place them in whatever order—S Adv V ~ Adv S V, S O V ~ O S V etc., there are no formal markers indicating which category is extraclausal. A better and more natural solution is to conclude that the entire perspective set out by a cartographic revision of V2 is misleading. 'Partial' or 'residual' V2 languages do not exist, non-restrictive word order systems with verb movement to clause-internal positions cannot be extensions of restrictive V2 or V1/V2 systems.

4. Verb movement and left periphery in non-V2 languages

In this section we briefly examine two languages which are definitely not V2 in the classic sense (that we share and advocate for here), but still have some properties of 'non-strict' / 'residual' V2. The aim of the discussion below is to show that clause-internal verb movement *per se* is not sufficient to produce the whole range of phenomena associated with V2 and can actually give rise to quite different and sometimes very peculiar systems.

4.1. Ossetic

The enigmatic clause structure in Ossetic has been a challenge for many formal linguists who attempted to make it fit into the system based on more or less reasonable assumptions about what functional projection constitute the clause and how they are placed with respect to each other (cf. Lyutikova&Tatevosov 2009, Ershler&Volk 2009, Gareyshina et al. 2011, Ershler 2012a,b, Belyaev 2013, 2014). The most striking characteristic of the Ossetian clause is that constituents normally found at the left edge of the clause (that is, complementizers and *wh*-phrases) are located in the preverbal position, that is, clause-internally. Together with negative particle and negative XPs, they form a rigid preverbal complex that cannot be separated from the verb by any argumental XP. At the same time, the linear order of other constituents of the clause is—at least superficially—free, so that they can precede or follow the verb giving

rise to various IS-interpretations. Thus, in (14) a complex sentence embedding finite complement clause is demonstrated; note that the complementizer *kæj* 'that' can only occur preverbally, whereas argument XPs can be positioned to the left or to the right of the complementizer and verb.

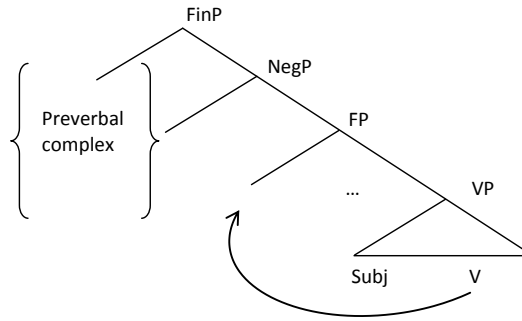
(14) a. æž žon-ın
I know-PRS.1SG
madinæ jæ fırt-ı **kæj** **arvıšt-a** goræt-mæ.
M. 3SG.GEN son-ACC that send.PST-TR.3SG city-LAT
'I know that Madina sent her son to the city.'
b. ...madinæ jæ fırtı gorætmæ **kæj arvıšta**.
c. ...**kæj arvıšta** madinæ jæ fırtı gorætmæ.
d. *... **kæj** madinæ jæ fırtı gorætmæ **arvıšta**.
e. *... **arvıšta** madinæ **kæj** jæ fırtı gorætmæ.

(15a–b) show the preverbal complex consisting of an interrogative *wh*-XP and a negative XP; crucially, the subject XP occupies different positions wrt the finite verb: it is strictly adjacent to the verb when being a negative pronoun and precedes the negative oblique complement when being an interrogative pronoun. The preverbal position of the subject XP in (15a–b) is thus a result of its movement to the preverbal complex in virtue of belonging to the category of NPI/*wh*-XPs, and not in virtue of being a subject. This reasoning is further supported by (15c) where the regular subject XP occurs postverbally.

(15) a. æž žon-ın kæj-mæ ničі azırd-ta.
I know-PRS.1SG who-COMIT nobody speak-PST.3SG
'I know whom nobody spoke to'.
b. æž žon-ın čі nikæj-mæ azırd-ta.
I know-PRS.1SG who nobody-COMIT speak-PST.3SG
'I know who spoke to nobody'.
c. æž žon-ın kæj nikæj-mæ azırd-ta zalinæ.
I know-PRS.1SG that nobody-COMIT speak-PST.3SG Z.
'I know that Zalina spoke to nobody'.

It follows from (14) and (15) that the verb must undergo a clause-internal movement in order to reach the structural position adjacent to the preverbal complex, thus leaving behind all the constituents c-commanding it in the VP and surfacing in some higher functional projection FP, as in (16). In Lyutikova & Tatevosov (2009 and elsewhere) it is argued that this functional projection is the T(ense)P, but nothing in our current argumentation crucially depends on this particular assumption. Whatever projection the verb moves in Ossetic, it shall contain no [EPP] feature, which otherwise would attract the dedicated XP (i.e. the subject of the clause if FP is indeed the TP) and create an intervenor between the verb and the preverbal complex.

(16)



To sum up, the clausal architecture of Ossetic calls for clause-internal verb movement and against grammatical feature-driven subject movement, thus {-EPP} in Roberts' (2012) taxonomy. At the same time, Ossetic is hardly a V2 language in the classic sense, as it does not meet the crucial 'bottleneck' condition for (V1/)V2. Needless to say that verb movement does not obligatorily result in complementary distribution of finite verbs and subordinators, even if they show strong positional interactions.

4.2. Russian

Russian is an example of a language where word order patterns sometimes mimic the true V2 languages' template. Thus, at least since Kovtunova (1976) the pairs of sentences like (17a–b) are considered as derivationally related, so that the preverbal position is occupied by the topical constituent, and the verb forms a rhematic (wide focus) IS-constituent with the postverbal material. Elena Paducheva (2008 and elsewhere) dubs the (b) examples as involving the subject inversion that she considers as a postsyntactic LA-transformation (examples are from Paducheva 2008).

(17) a. [$_{Th}$ Lodka] [$_{Rh}$ ležala na beregu].
'The boat lay on the shore.'
b. [$_{Th}$ Na beregu] [$_{Rh}$ ležala lodka].
'On the shore, there was a boat lying.'

John Bailyn (2012) attempts to provide an intra-syntactic account to the subject inversion; he claims that the preverbal constituent in both (17a) and (17b) (and similar examples) is in the structural subject position, i.e. Spec, TP, and the verb has moved to T. Thus, he treats the subject inversion separately from other word order permutations motivated by information structure. If we follow Bailyn's analysis of Russian "generalized inversion" constructions, we have to admit that the verb undergo the clause-internal movement out of the VP in order to precede the subject, and the specifier position of the target projection is not allocated to the subject exclusively, that is, Russian is {-EPP} in Roberts' (2012) taxonomy. Despite of these properties of Russian, however, we are not inclined to tag it as a 'partial V2' language; and we strongly doubt that other 'partial' / 'residual' V2 languages with similar characteristics exist.

5. Conclusions

In this paper, we have discussed the characteristic properties of V2 languages and the parameters of variation attested among them. This parametric representation of the complex V2 phenomenon can constitute an empirical basis of a theoretical work aiming at modelling the clausal architecture, as well as be implemented in rule-based parsing systems making use of adjustable parameter setting.

We have paid special attention to the recent claim that the broader definitions of V2 phenomena are possible which rely on a single parameter of the clause-internal verb movement and abandon other conditions on the clause architecture. We have shown that V2 is a complex phenomenon that cannot be reduced to the verb movement. Hence, non-restrictive word order systems disregarding the 'bottleneck' condition cannot be an extension of restrictive V2 and V1/V2 languages.

Abbreviations

1/2/3–1st/2nd/3rd person, ACC—accusative, Adv—adverbial, AdvP—adverbial phrase, AUX—auxiliary, CL—clitic, COMIT—comitative, Comp—complementizer, CP—complementizer phrase, DAT—dative, DEF—definite, DP—determiner phrase, ERG—ergative, FinP—finiteness phrase, FUT—future, GEN—genitive, INF—infinitive, IS—information structure, LA—linear-accent transformations, LAT—lative, M/F/N—masculine/feminine/neuter, NEG—negation, NOM—nominative, NP—noun phrase, O—object, PERF—perfective, PL—plural, PP—prepositional phrase, PRS—present, PRT—preterite, PST—past, S—subject, SG—singular, TP—tense phrase, TR—transitive, V—verb, V1—verb in the 1st position, V2—verb in the 2nd position, V_{fin}—finite verb, VP—verbal phrase, *wh*—interrogative pronoun.

References

1. *Bailyn, John Frederick.* (2012). The Syntax of Russian. Cambridge, Cambridge University Press.
2. *Bech, Kristin & Salvesen, Christine Meklenborg* (2014). Preverbal word order in Old English and Old French. In: Information Structure and Syntactic Change in Germanic and Romance Languages. Pp. 233–269.
3. *Belyaev, Oleg* (2013). Superiority effects in Ossetic: structural vs. linear constraints // (Typology of Morphosyntactic Parameters, 15–17 окт. 2013). — Moscow.
4. *Belyaev, Oleg* (2014). Korrelyativnaya konstrukciya v osetinskom yazyke v tipologicheskom osveshchenii [Ossetian correlative constriuction in the typological perspective]. (In Russian). PhD thesis, Moscow State University.
5. *Beninca, Paola & Poletto, Cecilia* (2004). "Topic, Focus, and V2". In: L. Rizzi (ed.). The Structure of CP and IP: Oxford Studies in Comparative Syntax. Oxford. Pp. 52–75.
6. *Bhatt, Rajesh M.* (1999). Verb movement and the syntax of Kashmiri. Dordrecht.

7. *Den Besten, Hans* (1983). "On the interaction of root transformations and lexical deletive rules". // On the formal nature of the Westgermania /Abraham W. (ed.). John Benjamins, Amsterdam, pp. 47–131.
8. *Diderichsen, Poul* (1976). Elementary Danish grammar [Elementær Dansk Grammatik]. (In Danish). København, 3rd edition ([1]1946).
9. *Erschler, David* (2012a). "Sluicing-like Phenomena and the Location of CP in Ossetic" // Proceedings of IATL28 (Tel Aviv University, 16–17 окт. 2012).—https://www.academia.edu/2494508/Sluicing-like_Phenomena_and_the_Location_of_CP_in_Ossetic
10. *Erschler, David* (2012b). "From preverbal focus to preverbal "left periphery": The Ossetic clause architecture in areal and diachronic perspective" // Lingua, Vol. 122, № 6. Pp. 673–699.
11. *Ershler, David & Volk, Vitaly* (2009). Puzzles of Digor Ossetic Negation // International conference on iranian languages (ICIL3), September 11–13, 2009, Paris.
12. *Faarlund, Jan Terje, Lie, Svein, and Vannebo, Kjell Ivar* (1997). The reference grammar of Norwegian [Norsk referansegrammatikk]. (In Norwegian). Universitetsforlaget, Oslo.
13. *Gareyshina, Anastasiya, Grashchenkov, Pavel, Lyutikova, Ekaterina, and Tatevosov, Sergei* (2011). Sentential proforms and complementation in Ossetian // 33 Jahrestagung der Deutschen Gesellschaft für Sprachwissenschaft. Text: Strukturen und Verarbeitung (Georg-August-Universität Göttingen, 23–25 Feb. 2011). [URL: http://www.zas.gwz-berlin.de/fileadmin/veranstaltung_zas/workshops/archiv/proforms/gareyshina_ossetian.pdf]
14. *Holmberg, Anders & Platzack, Christer.* (1995). The Role of Inflection in Scandinavian Syntax. Oxford.
15. *Holmberg, Anders* (2015). "Verb second". // Syntax—Theory and Analysis. An International Handbook. T. Kiss & A. Alexiadou (eds.). Vol. 1. Berlin: Mouton de Gruyter, pp. 242–283.
16. *Kaiser, Georg & Zimmermann, Michael* (2015). "Exploring verb order differences between (Medieval) Romance and (Modern) Germanic". Traces of History Conference. The University of Oslo, March 9–10, 2015.
17. *Kaiser, Georg* (2002). Verb position and its change in Romance. [Verbstellung und Verbstellungswandel in den romanischen Sprachen]. (In German). Tübingen: Niemeyer.
18. *Kovtunova, Irina* (1976). Russian syntax: word order and information structure [Russkiy sintaksis: porjadok slov i aktual'noye chleneniye predlozheniya]. (In Russian). Moskva: "Prosveshcheniye".
19. *Lyutikova, Ekaterina & Tatevosov, Sergei* (2009). The clause internal left edge: Exploring the preverbal position in Ossetian // International conference on iranian languages (ICIL3), September 11–13, 2009, Paris.
20. *Mathieu, Eric* (2006). "Stylistic Fronting in Old French". Probus, 18. Pp. 219–266.
21. *Paducheva, Elena* (2008). "Categoricity and ways to overcome it: subject inversion" [Kommunikativnaya raschlenennost' i puti ee preodoleniya: inversiya podlezhashchego] (In Russian) // Phonetics and Non-phonetics: To Sandro V. Kodzasov on the occasion of his 70th anniversary [Fonetika i nefonetika.

K 70-letiyu Sandro V. Kodzasova]. A. Arkhipov et al. (eds.). Moskva, "Yazyki slavyanskikh kul'tur". — P. 417–426 .
22. *Platzack, Christer* (2008). The Edge Feature on C. Lund, University of Lund, Ms.
23. *Rizzi, Luigi* (1997). "The Fine Structure of the Left Periphery". // Elements of Grammar: Handbook of Generative Syntax. Dordrecht. Pp. 281–337.
24. *Roberts, Ian* (2012). "Phases, Head-movement and Second-Position Effects". // Phases developing the framework / Angel J. Gallego (ed.). Berlin, Boston: Mouton de Gruyter, pp. 385–440.
25. *Salvesen, Christine Meklenborg* (2013). "Topics and the Left Periphery. A comparison of Old French and Modern Germanic". // In search of universal grammar: from Old Norse to Zoque. Pp. 131–172.
26. *Vikner, Sten* (1995). Verb Movement and Expletive Subjects in the Germanic Languages. Oxford Studies in Comparative Syntax, New York, Oxford: Oxford University Press.
27. *Wolfe, Sam* (2015). A New Perspective on V2 and the Evolution of Romance Clausal Structure // Traces of History Conference. The University of Oslo, March 9–10, 2015.
28. *Yanko, Tatiana* (2001). Communicative strategies of Russian speech. [Kommunikativnye strategii russkoj reci] (In Russian). Moscow
29. *Zimmerling, Anton* (2002). Typological Scandinavian syntax. [Typologicheskiy sintaksis skandinavskikh yazykov] (In Russian). Moscow.
30. *Zimmerling, Anton* (2013). Slavic word order systems from the viewpoint of formal typology. [Sistemy poryadka slov slavyanskikh yazykov v tipologicheskom aspekte] (In Russian). Moscow.
31. *Zimmerling, Anton & Kosta, Peter* (2013). "Slavic clitics: a typology". STUF—Language Typology and Universals. Vol. 66. No 2. Pp. 178–214.

Abstracts

CORRELATION BETWEEN SEMANTIC AND COMMUNICATIVE PROPERTIES OF WORDS

Apresjan V. Ju. (valentina.apresjan@gmail.com), National Research University Higher School of Economics; Vinogradov Russian Language Institute of the Russian Academy of Sciences, Moscow, Russia

The objective of this paper is to determine what semantic components in the meaning of a word facilitate its lexicalization as prosodically marked and aid its focalization in an utterance. The paper demonstrates that prosodic and communicative properties of a word correlate with its semantic properties. In particular, a case study of different senses of the words *tol'ko* 'only', *pravda* 'true', *eshche* 'still, more', *voobshche* 'in principle, generally', *po krajnej mere* 'at least' and some others reveals that focalization and prosodic marking in a word are triggered by the semantics of contrast, high degree, and addition. On the other hand, semantics of concession in the meaning of a word limits its ability for accentual marking and focalization. The observed correlations between semantics/on the one hand, and prosody and communicative properties, on the other, are confirmed by the multimedia corpus data.

JUSTICE VERSUS INJUSTICE: METAPHORICAL INTERPRETATIONS IN MODERN RUSSIAN DISCOURSE (THROUGH TEXTCORPUS OF PRINT MEDIA)

Baranov A. N. (Baranov_anatoly@hotmail.com), Institute of Russian Language RAS (Vinogradov's intstitute), Moscow, Russia

In the paper words *spravedlivost'* (justice) and *nespravedlivost'* (injustice) in Russian and their corresponding concepts are considered. It is shown that formally words *spravedlivost'* and *nespravedlivost'* are antonyms, because morphologically they differ only in morpheme *ne-* ("no"). But their meanings differ in a more complicated way. Word *spravedlivost'* has an abstract meaning, it denotes a value category. At the meantime extensional set of the word *nespravedlivost'* is another one: it is used for denoting of wide range of situations where features of justice as a value concept are violated. For this reason words *spravedlivost'* de facto is singularia tantum: it has not plural. At the same time the word *nespravedlivost'* (injustice) has in Russian speech bide forms: singular, as well as plural.

Differences in semantics between two words under consideration become apparent in metaphorical models which are used by speakers in interpretation of justice an injustice in Russian public discourse, model of which is text corpus of print media.

AUTOMATIC IDENTIFICATION OF SHARED ARGUMENTS IN VERBAL COORDINATIONS

Aleksandrs Berdičevskis (aleksandrs.berdicevskis@uit.no), **Hanne Eckhoff** (hanne.m.eckhoff@uit.no), UiT The Arctic University of Norway, Tromsø, Norway

We describe automatic conversion of the SynTagRus dependency treebank of Russian to the PROIEL format (with the ultimate purpose of obtaining a single-format diachronic treebank spanning more than a thousand years), focusing on analysis of shared arguments in verbal coordinations. Whether arguments are shared or private is not marked in the SynTagRus native format, but the PROIEL format indicates sharing by means of secondary dependencies. In order to recover missing information and insert secondary dependencies into the converted SynTagRus, we create a simple guessing algorithm based on four probabilistic features: how likely a given argument type is to be shared; how likely an argument in a given position is to be shared; how likely a given verb is to have a given argument; how likely a given verb is to have a given argument frame. Boosted with a few deterministic rules and trained on a small manually annotated sample (346 sentences), the guesser very successfully inserts shared sub-

jects (F-score 0.97), which results in excellent overall performance (F-score 0.92). Non-subject arguments are shared much more rarely, and for them the results are poorer (0.31 for objects; 0.22 for obliques). We show, however, that there are strong reasons to believe that performance can be increased if a larger training sample is used and the guesser gets to see enough positive examples. Apart from describing a useful practical solution, the paper also provides quantitative data about and offers non-trivial insights into Russian verbal coordination.

MARKERS OF WORD PRODUCTION DIFFICULTIES IN NORMAL AND CLINICAL DISCOURSE PRODUCTION: CONTINUITY OF NORM IN LANGUAGE AND DISCOURSE

Bergelson M. B.[1], **Akinina Yu. S.**[1], **Dragoy O. V.**[1], **Iskra E. V.**[1,2], **Khudyakova M. V.**[1],
[1]National Research University Higher School of Economics, Moscow, Russia;
[2]Center for Speech Pathology and Neurorehabilitation, Moscow, Russia

Aphasia is language impairment due to brain damage. Word-finding and word-retrieval problems can be very prominent in the speech of people with aphasia, being detectable in almost every aphasic speaker. On the other hand, word-finding difficulties and speech errors can sometimes occur in speech of neurologically healthy people. It is assumed that the same psycholinguistic levels of word-retrieval breakdown can account for the mistakes of both groups. In the meanwhile, retrieving of a single word from mental lexicon is not the only possible level of hindrance for a speaker: referential and lexical choices that take place at more general discourse and pragmatic level can also be disturbed.

The Russian CLiPS—Russian CLinical Pear Stories—is a corpus of film-elicited narratives retrieved following (Chafe, 1980) methodology from healthy and language-impaired cohorts. The aim of our research was to investigate the characteristics of formal markers of word retrieval difficulties in narratives of neurologically healthy people and people with aphasia. Three types of markers were considered (discourse markers, false starts and self-corrections) in the nominations of common referents of Pear stories narratives. The markers at different breakdown levels are qualitatively analysed, creating a platform for future analysis.

THE CASE OF RUSSIAN SUBJECT PRO IN MACHINE TRANSLATION SYSTEM

Bogdanov A. V. (abogdanov@abbyy.com)[1], **Gorbunova I. M.** (igorbunova@abbyy.com)[1,2],
[1]ABBYY, Moscow, Russia, [2]Russian State University for the Humanities, Moscow, Russia

This paper concerns a problem of Russian floating quantifiers (also known as semipredicatives) in machine translation. Floating quantifiers in Russian (such as *оба* 'both', *один* 'alone', *сам* 'on one's own' etc) are inclined for case, number and gender and agree in those categories with the subject of the minimal (finite) clause containing them. However, the case of a floating quantifier in an infinitive clause varies according to the type of PRO control applied and some other structural characteristics of the infinitive clause. This poses a problem for rule-based machine translation, to choose the correct case for the quantifier at synthesis, or to link it correctly to its antecedent at analysis. A model-based machine translation system, such as ABBYY Compreno, can handle the case choice problem, as this paper is to show.

SEMANTIC ANALYSIS AND QUESTION ANSWERING: A SYSTEM UNDER DEVELOPMENT

Igor Boguslavsky, Vyacheslav Dikonov, Leonid Iomdin, Alexander Lazursky, Victor Sizov, Svetlana Timoshenko, A. A. Kharkevich Institute for Information Transmission Problems, Russian Academy of Sciences, Moscow, Russia

The paper presents a system of semantic analysis and a question answering system implemented on its basis for a specific subject domain: (European) football match news. As input, the system obtains a natural language question (in Russian), which it answers with an element (or elements) from the repository of individuals. The core part of the system is the semantic analyzer of natural language texts. For each sentence of the text processed, the special semantic analysis component of ETAP-3 linguistic processor constructs a semantic structure, which consists of a set of triples of the type **`semantic_relation (individual,individual)`**. Semantic relations and individuals consti-

tuting this structure correspond to the elements of the ontology, which can thus be viewed as a functional analogue of a dictionary for the semantic language. Semantic structures of sentences belonging to a particular text are integrated thanks to coreference and anaphora resolution and converted into an OWL-document, which is later used as a database. This database is supplemented by background knowledge from the repository of individuals concerning specific teams, football players, and games. Thanks to this resource, we are able to find an answer to the question using not only the data contained in different sentences of the text but also in the repository of individuals. If the user asks "*What team defeated the champion of Spain?"* while we have a text reporting that "*Slutsky's players outplayed Atletico Madrid*" then the system will establish the correspondence with the question, the text, and the records in the depository of individuals, and will come with the correct answer "CSKA". The semantic structure obtained from the natural language question is converted into a SPARQL query addressed to the database. Currently, all parts of the system are operating in the test mode.

QUANTITATIVE METHODS IN DIACHRONIC LINGUISTIC STUDIES: THE CASE OF RUSSIAN DATIVE SUBJECTS WITH PREDICATIVES

Anastasia Bonch-Osmolovskaya (abonch@gmail.com), National Research University Higher School of Economics, Moscow Russia

The paper aims to demonstrate how quantitative corpus methods used in linguistics research may help to range different realizations of the same phenomenon: the use of dative subjects in predicative and adjective constructions. The core idea of the research is to study the distribution of dative subject constructions with predicative and adjective forms that potentially can be used in such constructions, i.e. aptitude of the construction for explication or omitting the dative subject. While usually the predicates are classified on the basis whether they can potentially be used with dative subject, I study the trends for explicit use of dative (or prepositional beneficiary arguments) among the "dative subject predicates", and show that the frequency rates of real use of dative subjects can be very different with different predicates. Separate analysis of different morphological forms of the same dative subject lexeme (i.e. adjectives in full and short forms, comparative adjectives and predicatives) shows that they may also exhibit different strategies with explicit dative subjects. Finally data from the 18th and the 21st centuries is compared and hierarchical clustering is used to reveal some diachronic trends.

STEM INITIAL ALTERNATION IN RUSSIAN THIRD PERSON PRONOUNS: VARIATION IN GRAMMAR

Daniel M. A. (misha.daniel@gmail.com), National Research University Higher School of Economics; Moscow State University; University of Helsinki

The paper discusses the present stage of the evolution of the initial [n]/[j] stem alternation in Russian third person pronouns. After providing a short overview of the origins of the forms, I focus on their category status, discuss Zalizniak's 'adpositionality' in some detail, and then proceed to considering the cases where the 'n'-forms are induced by a distant 'controller'. I will show that the fact that the 'n'-forms are essentially variants is better accounted for by the notion of 'trigger' of a morphological variant. To my eyes, this opens ways to a better understanding of the observed evidence than that using the conventional notion of morphosyntactic controller, on the one hand—and certainly than explaining them in (morpho)phonological terms. In the end, I will briefly argue that, in a sense, the evolution of the alternation is similar to degrammaticalization, showing a movement from a morphophonologically conditioned external sandhi to a morphosyntactic category similar to government.

MODAL PARTICLES AND THE ACTUALIZATION OF FORGOTTEN DETAILS (BASED ON THE MATERIALS OF PARALLEL CORPORA)

Dmitrij Dobrovol'skij (dm-dbrv@yandex.ru), **Irina Levontina** (irina.levontina@mail.ru), RLI RAS, Moscow, Russia

The use of parallel corpora carries with it special problems, particularly when it comes to units that are typical of oral speech. Nevertheless, it is the presence of good Russian-English and Russian-German parallel texts in the RNC that has made the present study possible. Our analysis also demonstrates the limitations inherent in investigations based on parallel corpora, especially with respect to discursive words. Only a combination of various research methods is capable of producing adequate results.

In this study we analyze a group of discursive words whose semantics actualize things that have been forgotten. There are two types of situations here. In the first, the speaker reminds the addressee of some object or event; in the second, the speaker attempts to remember some detail or name connected with the events s/he is talking about. Certain discursive words apply to both types of contexts, whereas others are special to one of them. Among the Russian discursive words that express these ideas are units such as *biš', tam, ešče, pomnite, ėtot* and phrasemes like *kak tam, kak ego, kak že, ėto samoe*, etc. We also examine English and German equivalents used to translate these words.

Russian has a rich repertory of discursive resources for actualizing forgotten details. In English, if the corresponding meanings are expressed at all, it tends to be done either syntactically or by means of explicit utterances. The German arsenal of discursive resources is no less extensive than the Russian, but there are no one-to-one correspondences between the Russian and German discursive words. They have different semantic configurations, and although the meaning components are often quite similar, they combine differently, so that translations of such particles in various contexts are rather diverse.

SUBJUNCTIVE PARTICLE AS A PART OF CONJUNCTION

Dobrushina N. R. (nina.dobrushina@gmail.com), National Research University Higher School of Economics, Moscow, Russia

Russian subjunctive is expressed by an analytical form which consists of subjunctive particle *by (b)* and past indicative or infinitive or a few predicative adverbs and adjectives. The subjunctive particle is an enclitic. It often merges with subordinate conjunctions, which yields words functioning as conjunctions and containing the subjunctive particle. Historically, the particle *by* in conjunctions can be traced back to the marker of subjunctive. Synchronically, however, the group is not homogenous. The aim of the paper is to find out which of the conjunctions with *by* should be considered as containing the marker of subjunctive, and test whether the particle can or can not be separated from the conjunction. Four criteria are used. The first and the second, namely, (a) the forms available in the subordinate clause with the conjunction and (b) the possibility of repetition of the particle *by* with the second predicate shows that comparative conjunctions do not synchronically contain the subjunctive marker. The third and fourth criteria, namely (c) the omission of the particle *by* and (d) its ability to be separated from the conjunction by another words give different results.

REFERENT INTRODUCTION IN RUSSIAN SPOKEN NARRATIVES

Fedorova O. V. (olga.fedorova@msu.ru), Lomonosov Moscow State University, Moscow, Russia

In a series of papers published twenty years ago on analysis of Russian, German and Shan tales, we examined the typology of referent introduction in written texts. The general purpose of the current study was to evaluate how the "tale" model of introduction is applicable to the spoken narratives; as an alternative approach we considered the Chafe's model, based on the English "Pear stories" (Chafe 1980). Twenty five Russian participants took part in the experiment; all the participants described the same experimental film about some child stealing pears; thus we analyzed 25 narratives and 125 introductive sentences. Surprisingly, our model differs from both the "tale" introductive model and the Chafe's model for each of the following points: (1) type of the common ground, (2) speech disfluencies, (3) the character status and clauses number, (4) the "light subject" constraint, (5) the "one new idea" constraint. However, all of these results need further empirical justification in new studies on Russian materials.

DOCUMENT VS. META-DOCUMENT: ARE THEIR RHETORIC STRUCTURES DIFFERENT?

Galitsky B. A. (bgalitsky@hotmail.com), Knowledge-Trail Inc. San Jose CA USA

The problem of classifying text with respect to belonging to a document or a meta-document (metalanguage and language object patterns) is formulated and its application areas are proposed. An algorithm is proposed for document classification tasks where counts of words is insufficient do differentiate between such abstract classes of text as metalanguage and object-level. We extend the parse tree kernel method from the level of individual sentences towards the level of paragraphs, based on anaphora, rhetoric structure relations and communicative actions

linking phrases in different sentences. Tree kernel learning is then applied to these extended trees to leverage of additional discourse-related information. We evaluate our approach in the domain of action-plan documents, as well as in literature domain, recognizing some portions of text in Kafka's novel "The Trial" as metalanguage patterns and differentiating them from the novel's description in the studies of Kafka by others.

LINGUISTIC ANALYSIS IN THE SPEAKER IDENTIFICATION SYSTEMS: INTEGRATED COMPLEX EXAMINATION APPROACH BASED ON FORENSIC SCIENCE TECHNOLOGY

Galyashina E. I. (galyashina@gmail.com), Kutafin Moscow State Law University, Moscow, Russia

The article proposes the concept of the integrated expert techniques for speaker identification on the domain of complex acoustic-phonetic and linguistic methods of oral speech analysis, defining general and special forensic expert competences. The result of forensic speaker identification used as evidence must exhibit a high level of reliability. The author examines the key concepts and terms of the procedure of individual-specific speaker identification in the aspect of modern expertology (forensic science). The paper states the need to take into account that the role of professional linguistic competences increases in conditions when digitized speech signals are compared, algorithm of coding is indefinite and falsification of utterances is not excluded. To solve this problem the author proposes a multistage approach consisting of a parallel application of instrument and technical methods together with aural-perceptual, waveform and sonogram investigation and sophisticated linguistic analysis. The main attention is paid to the linguistic component of the complex integrated approach based on the phonetic and semantic analyses. It is stated that individualized speech unit is formed by a system of miscellaneous formal and semantic relations of structural speech components in linguistic contents. The proposed method of integration of the multilevel speech modules was implemented in forensic linguistic methodology of speaker identification technique. This made it possible to considerably increase the reliability of the expert's decision and provided an opportunity to use it as a component of the multistage system for speech utterances authentication.

MODEL-BASED WSA AS MEANS OF NEW LANGUAGE INTEGRATION INTO A MULTILINGUAL LEXICAL-SEMANTIC DATABASE WITH INTERLINGUA

Goncharova M. B. (maria_go@abbyy.com), **Kozlova E. A.** (Helen_Koz@abbyy.com), **Pasyukov A. V.** (Artem_P@abbyy.com), **Garashchuk R. V.** (Ruslan_G@abbyy.com), **Selegey V. P.** (Vladimir_S@abbyy.com), ABBYY, Moscow, Russia

This paper presents a model-based approach to Word Sense Alignment (WSA) applied for new language integration within ABBYY Compreno lexical-semantic database with interlingua. Using the model, i.e. semantic and syntactic compatibility, we perform semantic-syntactic analysis with language-independent structure as a result. With the comprehensive description of core languages at our disposal, we analyze parallel resources, namely, the part of a bilingual dictionary and of a parallel corpus in a source language, and obtain a set of candidate concepts for meanings of a target language. In this way, we accomplish WSA between the dictionary meanings and the concepts of interlingua. Once the correspondences between the meaning and the concepts of the hierarchy are established, these new meanings can be incorporated into the lexical-semantic database. The integration is fulfilled semi-automatically, i.e. at the final stage the correspondences are to be approved by a linguist; however, the amount of manual work is reduced to minimum.

QUANTIFIERS, GESTICULATION, AND VIEWPOINT

Grishina E. A. (rudi2007@yandex.ru), Vinogradov Institute of Russian Language RAS, Moscow, Russia

The study analyzes gestures, which regularly accompany Russian universal quantifiers *ves'* 'the whole of', *vse* 'all', *kazhdyy* 'every', *l'uboy* 'any'. The results of the study shows that the accompanying gesticulation correlates more with the pragmatic features of the quantifiers (the spatial and evidential speaker's position and the speaker's modus operandi), than with the logical components of the quantifier's semantic structure. The Multimodal Russian corpus (MURCO) has been used as a source of the data.

THE APPLICATION OF MACHINE LEARNING METHODS FOR ANALYSIS OF TEXT FORUMS FOR CREATING LEARNING OBJECTS

Grozin V. A. (vlad.grozin@yandex.ru), **Dobrenko N. V.** (graziokisa@gmail.com), **Gusarova N. F.** (natfed@list.ru), ITMO University, Saint Petersburg, Russia; **Ning Tao** (603096136@qq.com), Changchun University of Science and Technology, Changchun, China

Nowadays the concept of a learning object (LO) is widely used in preparation of educational materials. Usually, LOs are parts or fragments of previously created educational content, which is very informative and pedagogically focused. However, concerning high-dynamic branches of science and technologies LOs tend to become outdated and trivial thus losing their educative value. In this situation, specialized text forums become a valuable source of knowledge. Forums contain experience of people who actually used the technology and its features. They contain both positive and negative experience—something that is not available from official documentation at all. However, they also contain many trivial, repeated and still irrelevant posts. Also, an expert needs to extract useful messages from text forums according to his individual learning objectives.

The paper deals with the task of automatically identifying texts potentially useful for preparation of textual educational materials within text forums. For our experiments, we have selected highly inflective languages with complex grammar and rather weak text analysis tools: French, German, Russian and Chinese (Mandarin). We have overviewed non-semantic text and social features of a text forum which indicate the suitability for creation of a textual LO. We have analyzed those features. For this purpose, we have constructed linear and non-linear models of machine learning and conducted feature selection. Even for the forums providing little information about chosen topics and forums with a lot of off-topic text in dataset, these models were better than the baseline selection methods.

NUTS: WHAT ARE THEY?

Iomdin B. L. (iomdin@ruslang.ru), V. V. Vinogradov Russian Language Institute of the Russian Academy of Sciences, National Research University "Higher School of Economics", Moscow, Russia

When describing words which denote real life objects, dictionaries tend to use scientific terms and classifications, even when dealing with natural language. This approach may lead to misunderstanding, especially in cases when scientific classification (e. g. in biology) differs from what is found in natural language data. One of such cases is discussed here, namely the small but rather interesting class of nuts (Russian *orexi*). In the botanic world view nuts usually include hazelnuts and chestnuts, but do not include almonds (which are considered stone fruits), pine nuts (seeds), peanuts (legumes), pistachio (kernels), etc. The Russian *orex*, English *nut*, Latin *nux* exhibit similar behaviour here. Explanatory dictionaries of Russian more or less follow the botanical definitions, even though in many fields (such as cooking, food industry, medicine, etc.) nuts are classified differently. In order to establish the boundaries of nuts in Russian, more than 1,000 native speakers were questioned and multiple texts of different periods were studied. The result is a peculiar class which could not be identified with any of the natural language supercategories described by Anna Wierzbicka. A new lexicographic description is proposed for some words included into this class.

THE PROBLEM OF NON-DISCRETENESS AND SPOKEN DISCOURSE STRUCTURE

Kibrik A. A. (aakibrik@gmail.com), Institute of Linguistics RAS; Lomonosov Moscow State University, Moscow, Russia

Language consists of units of various hierarchical levels, but the boundaries between the units are not always crisp, and non-discrete effect are observed. That applies not only to syntagmatic structure, but also to paradigmatics, diachrony, and even whole languages. Non-discreteness is a common property of language and cognition. In contrast to conventional discrete and continuous structures, I propose another kind of structure that can be called focal. Focal phenomena are simultaneously distinct and related. It is necessary to recognize focal structure as one of the major types of structures typical of natural language. Non-discrete effects can be observed at the level of discourse. Spoken discourse consists of elementary discourse units (EDUs), identifi-

able with the help of a set of behavioral criteria. Along with prototypical clausal EDUs, there are deviant EDUs of various kinds. Parcellated elaborations constitute an example of a paradigmatic outlier among the EDUs. Non-discrete boundaries between EDUs are an illustration of syntagmatic difficulties in EDU identification. Phonemes, EDUs, and other units are not as crisp and clean as our digital mind would want them to be. In order to address linguistic reality in its actual complexity, we have to recognize that segmentation follows the principles of focal structure, which is the general property of language and cognition.

DEVELOPMENT OF FACTORED LANGUAGE MODELS FOR AUTOMATIC RUSSIAN SPEECH RECOGNITION

Kipyatkova I. S. (kipyatkova@iias.spb.su)[1,2], **Karpov A. A.** (karpov@iias.spb.su)[1,3]
[1]St. Petersburg Institute for Informatics and Automation of RAS (SPIIRAS), St. Petersburg, Russia
[2]St. Petersburg State University of Aerospace Instrumentation (SUAI), St. Petersburg, Russia
[3]ITMO University, St. Petersburg, Russia

In this paper, we present a study of factored language models (FLM) of Russian for rescoring N-best lists in automatic speech recognition (ASR) systems. We used 3-gram language models as baseline. Both 3-gram and factored language models were trained on a text corpus collected from recent Internet online newspapers; total size of the text corpus is about 350 million words (2.4 Gb data). For FLM creation, we used five linguistic factors: word-form, word lemma, stem, part-of-speech, and morphological tag. We studied several FLMs with two factors (word-form plus one of the other factors) using 2 fixed backoff paths: (1) the first drop was of the most distant word and factor, then—of the less distant ones; (2) the first drop was of the words in time-distance order, then drop of the factors in the same order. We investigated the influence of a factor set and backoff paths on language model perplexity and word error rate (WER). Also we created FLMs with some parallel generalized backoff paths. Optimization of the FLM parameters was carried out by means of the genetic algorithm. The FLMs were embedded in the automatic Russian speech recognition system with a very large vocabulary. Experimental results on continuous Russian speech recognition task showed a relative WER reduction of 8% when the FLM was interpolated with the baseline 3-gram model.

RUSSIAN LEXICOGRAPHIC LANDSCAPE: A TALE OF 12 DICTIONARIES

Yuri Kiselev (ykiselev.loky@gmail.com)[1, 2], **Andrew Krizhanovsky** (andrew.krizhanovsky@gmail.com)[3], **Pavel Braslavski** (pbras@yandex.ru)[1, 4], **Ilya Menshikov** (unkmas@gmail.com)[1], **Mikhail Mukhin** (mfly@sky.ru)[1], **Nataly Krizhanovskaya** (nataly@krc.karelia.ru)[3], [1]Ural Federal University, Ekaterinburg, Russia; [2]Yandex, Ekaterinburg, Russia; [3]Institute of Applied Mathematics Research, Karelian Research Center of RAS, Petrozavodsk, Russia; [4]Kontur Labs, Ekaterinburg, Russia

The paper reports on quantitative analysis of 12 Russian dictionaries at three levels: 1) headwords: the size and overlap of word lists, coverage of large corpora, and presence of neologisms; 2) synonyms: overlap of synsets in different dictionaries; 3) definitions: distribution of definition lengths and numbers of senses, as well as textual similarity of same-headword definitions in different dictionaries. The total amount of data in the study is 805,900 dictionary entries, 892,900 definitions, and 84,500 synsets. The study reveals multiple connections and mutual influences between dictionaries, uncovers differences in modern electronic vs. traditional printed resources, as well as suggests directions for development of new and improvement of existing lexical semantic resources.

NOTES ON THE STRUCTURE OF CIRCUMFIXAL VERBS IN RUSSIAN

Kisseleva X. L. (xenkis@mail.ru), Vinogradov Institute for Russian Languagee, Moscow, Russia; **Tatevosov S. G.** (tatevosov@gmail.com), Lomonosov Moscow State University and Higher School of Economics, Moscow, Russia

The paper argues for an analysis that reduces the derivation of non-compositional circumfixal verbs to a fully compositional combination of two pieces of morphology independently attested in Russian, a (resultative) prefix, and the reflexive morpheme *-sja*. Circumfical verbs are ana-

lyzed as involving the following steps of derivation. First, the activity event structure projected by a non-derived stem is augmented by the change-of-state component that turns it into an accomplishment even structure. Secondly, prefixation occurs that introduces the maximal degree of change along a relevant scale creating a transitive verb. Such a verb, however, is ill-formed since the change-of-state component has not been licensed via lexical insertion. To rescue the derivation, reflexivization is invoked, and the change-of-state subevent gets licensed through identification of its participant with the clausal subject. A wider theoretical implication of the analysis is that circumfixation, as a primitive type of affixation, is superfluous and is to be abandoned.

VOWEL REDUCTION AS AN INDICATOR OF ITS STRESS IN STANDARD MODERN RUSSIAN

Knyazev S. V. (svknia@gmail.com), Moscow State Lomonosov University, Higher School of Economics, Moscow, Russia

The phonetic unity of phonological word in Standard Modern Russian is governed to a large extend by phonological rules of vowel realization, which forbid the reduced vowel in a position of first pretonic syllable (with the exception of some clitical syntactic words — prepositions and particles). The paper deals with the instrumental study of phonetic markers of stress in compound (predominantly loan) disyllabic words in Standard Modern Russian (e. g. *stop-krán* 'emergency brake') as compared with non-compound native words of the same phonological structure (*stoptát'* 'tread down'). The paper states that compound disyllabic words under a phrase accent have both syllables stressed, stress being signalized by mid [o] // [e] vowels impossible in unstressed position in Standard Modern Russian. In non-compound native words unstressed [o] vowel in all types of phrase positions after "hard" consonants is displaced by [a] (unstressed [a] being about 10 percent longer than stressed [o]). Our data shows that in compound disyllabic words in a position with no tonal accent phonetically unstressed /o/ is realized by reduced [ə] (not standard [a]) vowel (being somewhat twice shorter than unstressed [a]). Thus, non-trivial [o] vowel reduction in compounds may serve as a phonetic cue of phonological stress which is shown up fully only under the tonal accent. Phonetically the units in question should be treated as a combination of two phonological words and phonetic data may be used as a ground for orthographic adaptation of loan word.

ELEMENTARY DISCOURSE UNITS IN SPOKEN MONOLOGUES: EVIDENCE FROM COMMUNICATIVE PROSODY

Korotaev N. A. (n_korotaev@hotmail.com), RSUH, RANEPA, Institute of Linguistics (Russian Academy of Sciences), Moscow, Russia

The paper addresses the issue of spoken discourse segmentation. Using the corpus "Stories of presents and skiing", I explore the concept of Elementary Discourse Unit (EDU) — a chunk of speech flow defined on both prosodic and syntactic grounds. I propose for a new procedure to establish EDUs' boundaries. Compared to previous studies, a communicative perspective is added. I introduce the notion of communicative prosodic constituent, as well as a typology of those. It is based on three oppositions: (i) topic vs. comment; (ii) completion vs. transitional-continuity; (iii) main line vs. parenthesis. These oppositions are defined independently of one another and provide for a six-fold classification.

Several remarks should be made here. First, comments and (optionally) topics are found not only in statements, but also in other illocutionary types — such as questions, directives, vocatives, and so on. Second, it is sometimes hard to distinguish between comment constituents that express transitional continuity properties and topic constituents. I show that in some cases, this distinction can be made even though the intonation patterns are quite similar. Third, parenthetic constituents may as well have internal topics and comments.

Next, EDUs' boundaries are re-defined as a subset of communicative prosodic constituents' boundaries. Comment constituents always imply EDU's boundaries, while a topic constituent needs a syntactical support to do so. Finally, I provide an analysis of communicative structure and EDUs boundaries in an excerpt from the corpus.

FUNCTIONAL ANALYSIS OF NON-VERBAL COMMUNICATIVE BEHAVIOR

Kotov A. A. (kotov@harpia.ru), **Zinina A. A.** (zinina_aa@nrcki.ru), National Research Center "Kurchatov Institute", Moscow, Russia

In this study we represent functional annotation of the Russian Emotional Corpus (REC). The annotation is appended to the regular annotation of eyes, eyebrows and hand movements with supplementary annotation for head and corpus movements. The annotation records communicative functions, where a movement is intended for a particular goal or can be understood as connected to a particular goal/stimulus by the addressee. We show that a particular function can be expressed by different patterns, utilizing facial expression and/or hand/body movements. Functional annotation is also used as a non-terminal symbol in a generative grammar to produce non-verbal behavioral patterns.

HUMAN BODY IN AN ORAL DIALOG: THE CORPORAL FEATURE "SIZE OF THE SOMATIC OBJECT"

Kreydlin G. E. (gekr@iitp.ru), **Khesed L. A.** (lidakhe@yandex.ru), RSUH, Moscow, Russia

The paper continues a series of works on multimodal oral communication, many of which were printed in the Proceedings of previous «Dialogs». The interplay of verbal and nonverbal, mainly corporal, Russian sign codes in everyday communication is explored within the framework of the featured approach. The latter is based on the concept and instruments of the semiotic conceptualization of the human body, i.e. the naive map of how Russians think and talk about the body and its parts, organs, corporal liquids, covers, etc. and how Russians use somatic objects in various types of gestures, postures, sign movements, and other body meaningful units.

The core of the semiotic conceptualization holds several sets such as those of somatic objects and their natural language names, the sets of corporal features, their values and names, etc. In this paper, we focus primarily on the series of features named «the size of the somatic object» and provide some results of their language and nonverbal semiotic analyses. Two basic kinds of the features discussed that we call *an absolute size* and *a relative size* are distinguished, and the meaning and usage of many Russian expressions which reflect absolute and relative sized are described. Also, some correlations between some verbal and nonverbal Russian sign units of size are singled out.

THE DEPTH OF PROSODIC BREAKS IN SPOKEN TEXT (EXPERIMENTAL DATA)

Krivnova O. F. (okrivnova@mail.ru), Lomonosov Moscow State University, Moscow, Russia

This paper deals with the problem of prosodic phrasing in a spoken text. The introductory section provides a brief description of the background, clarifies basic terms and explains the concept of prosodic break and word boundary strength. The second section contains a short analysis of the current state of research in this area of phrasal prosody, highlights the main directions of the modern fundamental studies and applications, notes their relevance and the need to expand their empirical base. The third section deals with issues related to the local markers of prosodic phrasing, their hierarchy and phonetic means of realization. Here are given the examples of prosodic labeling of poetic and prose texts in the original transcription of famous Russian linguists Scherba and Avanesov with equivalent transcripts using quantitative, graduated scale of prosodic indexes similar to the labeling scheme adopted in foreign prosodic studies. Particular attention is paid to discussion of A.Sanderman's study, which is the most thorough contemporary analysis of prosodic phrasing. The fourth section describes the aim, material, technique and results of of perceptual and instrumental analysis of the location and depth of prosodic breaks carried out by the author of this paper on the Russian material. It is shown that native speakers quite consistently determine the location and depth of prosodic breaks using a 5-point rating scale, but breaks with minimum indexes are clearly opposed to the other types on the probability of their perceptual detection. Correlation of perceptual breaks' evaluation with pause duration at word boundaries is also investigated. In conclusion the material, methods and results of the experimental studies discussed in this paper are compared, the current trends in the use of the data are highlighted, the prospects and challenges for further studies of prosodic phrasing in speech are outlined.

PARTICLES 'VOT' AND 'VON': THE MECHANISMS OF SECONDARY MEANINGS FORMATION ON THE BASIS OF DEICTIC VALUES

Krylova T. V. (ta-kr@yandex.ru), V. V. Vinogradov Russian Language Institute of the Russian Academy of Sciences

This article is devoted to consideration of the particles VOT and VON. It is another attempt to bring secondary meanings of VOT and VON from their deictic meanings. After analyzing derived meaning of these particles we found no parallelism in the structure of their polysemy. We suggested that this is due to differences in the localization of the object in their index meaning ('proximity to the speaker' VS. 'remoteness from the speaker'). Next, we made an attempt to trace how these components are transformed in a secondary meanings of these words. We found that the component 'proximity to the speaker', which is included in the sense of VON, is transformed into component 'proximity to the moment of speech'. The last one passes in its transformation following stages (each the next is characterized by increasing of metaphoricalness): 'proximity to the moment of speech' → 'temporal proximity of the events' → 'interdependency of events'. Meanwhile the component 'remoteness from the speaker' included in the meaning of VON is transformed into component 'high degree' or into the indication of a reference to some fact (commonly known to speaker). Thus we can conclude that the asymmetry of VOT and VON derived values is due to the different direction of metaphorization of deictic components 'proximity to the speaker' VS. 'remoteness from the speaker'.

AUTOMATIC UPDATE OF THE NAMED ENTITIES DATABASE BASED ON THE USERS QUERIES

Kudinov M. (m.kudinov@samsung.com), **Piontkovskaya I.** (p.irina@samsung.com), Samsung R&D Institute, Moscow, Russia

We describe an algorithm of update of the database of named entities providing support of voice commands on a device. The update is made automatically with no human assistance by means of analysis of query logs of the dialogue system. The logs consisted of responses of the automatic speech recognition engine and thus contained erroneous recognitions. The search of such mistakes is also made as a part of our method. The problem of named entities extraction was solved by means of an algorithm based on entropy and mutual information statistics. The detection of recognition mistakes was made by means of a novel data-driven probabilistic approach taking into account grapheme substitution statistics in the data. Assuming grapheme alignment hidden, we use the EM algorithm for training the model. As a result we obtain a statistical model capable for sequence similarity assessment. The algorithm based on our similarity score performs better in terms of F_1-measure than one using the classical Levenshtein distance.

ABSTRACT LEXEMES' VALENCIES: REDUCTION VS. SPECIFICATION

Kustova G. I. (galinak03@gmail.com), Vinogradov Russian Language Institute of the Russian Academy of Sciences, Moscow, Russian Federation; National Research Tomsk State University, Tomsk, Russian Federation

Abstract vocabulary of different semantic classes (interpretation: *nepriyatnost'* ('nuisance', 'trouble', 'annoyance'), *promakh* ('a piece of carelessness'), *razlad* ('discord'); event: *referendum*, *soveshchanie* ('meeting'), *turnir* ('tournament'); activities: *mery* ('measures') / *rabota* ('work') / *usiliya* ('efforts') [*po uregulirovaniyu* ('on the settlement')] et al. are considered by analogy with speech and mental lexemes. The latter lexemes have valency on content (it is usually expressed by the subordinate clause) and valency on topic (it is usually expressed by the prepositional phrase *o X* 'about X').

Abstract lexemes valency on content / topic may also be expressed by the prepositional phrase *po X* [Dat.] 'on X': *mekhanizm po privlecheniyu klientov* 'mechanism to attract customers', *negativnaya tendentsiya po ukhudsheniyu portfel'a* ('negative trend for the deterioration of the portfolio'), *shagi po osushchestvleniyu mandata* ('steps to implement the mandate').

Valency on topic is both a reduction and a specification of the content of the situation.

AUTOMATIC DISAMBIGUATION IN THE CORPORA OF MODERN GREEK AND YIDDISH

Kuzmenko E. A. (eakuzmenko_2@edu.hse.ru), **Mustakimova E. G.** (egmustakimova_2@edu.hse.ru), National Research University Higher School of Economics, Moscow, Russia

The problem of morphological ambiguity is widely addressed in the modern NLP. Mostly ambiguity is resolved with the use of large manually-annotated corpora and machine learning. However, such methods are not always available, as good training data is not accessible for all languages. In this paper we present a method of disambiguation without gold standard corpora using several statistical models, namely, Brill algorithm (Brill 1995) and unambiguous n-grams from the automatically annotated corpus. All the methods were tested on the Corpus of Modern Greek and on the Corpus of Modern Yiddish.

As a result, more than a half of words with ambiguous analyses were disambiguated in both corpora, demonstrating high precision (>80%). Our method of morphological disambiguation demonstrates that it is possible to eliminate some of the ambiguous analyses in the corpus without specific linguistic resources, only with the use of raw data, where all possible morphological analyses for every word are indicated.

AUTOMATIC CLASSIFICATION OF WEB TEXTS USING FUNCTIONAL TEXT DIMENSIONS

Lagutin M. B. (lagutinmb@mail.ru)[1], **Katinskaya A. Y.** (a.katinsky@gmail.com)[2], **Selegey V. P.** (Vladimir_S@abbyy.com)[2,3,4], **Sharoff S.** (s.sharoff@leeds.ac.uk)[2,5], **Sorokin A. A.** (alexey.sorokin@list.ru)[1,2,3]
[1]Lomonosov Moscow State University, Moscow, Russia; [2]Russian State University of Humanities, Moscow, Russia; [3]Moscow Institute of Physics and Technology, Moscow, Russia; [4]ABBYY, Moscow, Russia; [5]Leeds University, Leeds, UK

The work addresses automatic genre classification of Web texts. We show that functional text dimensions could be used for this tasks, with their stable combinations (clusters) corresponding to genres. Basing on a gold standard corpus, we construct a list of such genres. We also show that functional dimensions values can be automatically extracted from language features. In the conclusion we discuss the application of our results for automatic annotation of large Web corpora.

AN EXPERIENCE OF CREATING MELODIC PORTRAITS OF COMPLEX DECLARATIVE SENTENCES OF RUSSIAN

Lobanov B. M. (Lobanov@newman.bas-net.by), United Institute of Informatics Problems of NAS Belarus, Minsk, Belarus

We proceed from the model of intonation patterns (IP) by Elena Bryzgunova, widely used in teaching Russian speech intonation. Bryzgunova distinguishes seven major Russian intonation patterns, named IP 1 to IP 7, of which only IP 1 is clearly used in declarative sentences to mark their completeness. The remaining six IPs are implemented for interrogative (IP2 — IP 4) or exclamatory (IP5 — IP7) types of sentences. Obviously, declarative sentences are overwhelming in professional and literary texts, particularly in professionally voiced texts of various genres (audio books). Most of them are not simple sentences and often consist of a mixture of complex and compound sentences.

The present study continues the author's paper “Universal melodic intonation portraits of Russian speech” presented to Dialogue 2014 conference, which introduced the concept of Universal Melodic Portrait (UMP). The present paper experimentally studies the intonation features of declarative sentences. It describes the results of auditory analysis and IP interpretation for declarative sentences of varying degrees of complexity, voiced by 3 speakers, and provides experimental representations of their intonation structures in the form of a sequence of Universal Melodic Portraits (UMP).

The paper is organized as follows. Section 1 describes the experimental procedure, including the characteristics of selected text and audio material, the listening method and the method of constructing a sequence of UMP’s audio recordings. Section 2 presents the experimental results: the graphical representation of an experimental sequence of universal melodic portraits of analyzed audio recordings. Section 3 offers an interpretation of the results.

INDUCING VERB CLASSES FROM FRAMES IN RUSSIAN: MORPHO-SYNTAX AND SEMANTIC ROLES

Olga Lyashevskaya (olesar@yandex.ru)[1,2], **Egor Kashkin** (egorkashkin@rambler.ru)[2]
[1]National Research University Higher School of Economics
[2]V. V. Vinogradov Russian Language Institute of RAS, Moscow

The paper presents clustering experiments on Russian verbs based on the statistical data drawn from the Russian FrameBank (framebank.ru). While lexicology has essentially abandoned the idea of syntactic transformations as the primary basis for grouping verbs into semantic classes (Apresjan 1967, Levin 1993), the hypothesis of the same lexical and syntactic distributional profiles underlying lexical clusters is still attractive. In computational linguistics, some attempts have been made to obtain verb classes for English, German and other languages using observable morpho-syntactic and lexical properties of context (Dorr and Jones 1996; Lapata 1999; Schulte im Walde 2006; Lenci 2014, among others).

Our experiments on semantic classification of Russian verbs are based on two types of tags embedded in the annotation of argument constructions: a) semantic roles and b) morpho-syntactic patterns. The domain of speech verbs is classified automatically on vectors, and the resulting clusters are contrasted against Babenko (2007)'s semantic classes and three other manual classifications. The classes within the domain of possessive verbs are constructed using rule-based solutions and evaluated against Berkeley FrameNet verb clusters. We conclude that clustering on morpho-syntactic (pure formal) patterns loses the race to more intelligent approaches which take into account semantic roles.

EXERCISE MAKER: AUTOMATIC LANGUAGE EXERCISE GENERATION

Malafeev A. Yu. (aumalafeev@hse.ru), National Research University Higher School of Economics, Nizhny Novgorod, Russia

Current trends in education, namely blended learning and computer-assisted language learning, underlie the growing interest to the task of automatically generating language exercises. Such automatic systems are especially in demand given the variability in language learning. Despite the abundance of resources for language learning, there is often a lack of specific exercises targeting a particular group of learners or ESP course. This paper gives an overview of a computer system called Exercise Maker that is aimed at flexible and versatile language exercise generation. The system supports seven exercise types, which can be generated from arbitrary passages written in English. Being able to tailor educational material to learners' interests is known to boost motivation in learners (Heilman et al., 2010). An important feature of the system is the automatic ranking of the source passages according to their complexity/readability. As shown by expert evaluation, the automatically generated exercises are of high quality: the gap precision is about 97–98%, while the overall exercise acceptance rate varies from 90% to 97.5%. Exercise Maker is freely available for educational and research purposes.

METALINGUISTIC PORTRAIT OF FASHIONABLE WORDS

Mustajoki A. (arto.mustajoki@helsinki.fi), University of Helsinki, Helsinki, Finland;
Vepreva I. T. (irina_vepreva@mail.ru), Ural Federation University, Ekaterinburg, Russia

The paper is based on a corpus study on the essential characteristics of fashionable word. The retrieval system for the identification of basic units consists of the utterances, containing meta-operators *modnoe slovo* (fashionable or trendy word) and *kak modno govorit'* (how to speak in a fashionably way). The theoretical model of fashion, worked out by A.B. Gofman, served as the basis for the interpretation of the findings. The work distinguishes context markers, manifesting attributive characteristics of the fashionable object, *modernity, universality, demonstrativeness* and *play*.

Based on the metalinguistic valuations corpus there were distinguished three classes of fashionable words, associated by commonplace consciousness with two attributive fashionable values—modernity and universality. These words include 1) the new words, naming the new reality; 2) the words, referring to the new naming of the known reality; 3) the only class of words meeting all the requirements of fashion, realizing play activity or speaker's aesthetic need to renew his speech.

The demonstrativeness markers are distinguished by the context indicators of the unusualness of the fashionable word form. In the first instance, the foreign neologisms demonstrate external attractiveness.

The class of the fashionable lexical items can be presented as a field structure with a nexus, having all the required value criteria of a fashionable object, and peripheral layers of various degree, depending on the number of criteria they have.

GRAPH-BASED APPROACH IN THE DEPENDENCY PARSING TASK FOR RUSSIAN LANGUAGE

Muzychka S. (s.muzychka@samsung.com), **Piontkovskaya I.** (p.irina@samsung.com), Samsung R&D Institute, Moscow, Russia

Dependency parsing is one of the key components in a large number of tasks of automatic processing of natural language texts. Effective dependency tree construction can be applied to a wide variety of machine translation systems, automatic speech synthesis and recognition, and so forth. Graph-based approach in dependency parsing proved to be efficient for morphologically rich languages due to its possibility to deal with non-projective dependency trees and flexible word order. Usually graph-based methods enable to perform probabilistic analysis over distribution on the set of syntax trees. In some NLP tasks it is not required to present a full syntactic parsing (in particular, to set labels on the edges of the tree). It is enough to find a parent for a given token. In this case, the graph-based approach is more appropriate because the likelihood that a token is an ancestor of the other, can be calculated by the explicit formula.

We consider a task of automatic syntax tree construction with application to Russian language corpus SynTagRus. We propose a novel technique which enables to reduce time costs for training and doesn't affect resulting accuracy. Experiments show that our algorithm outperforms existing analogues on SynTagRus in UAS (unlabeled attachment score) measure (percentage of correctly identified unmarked dependencies).

COREFERENCE CHAINS IN CZECH, ENGLISH AND RUSSIAN: PRELIMINARY FINDINGS

Anna Nedoluzhko (nedoluzko@ufal.mff.cuni.cz)[1], **Svetlana Toldova** (stoldova@hse.ru)[2], **Michal Novák** (mnovak@ufal.mff.cuni.cz)[1]
[1]Charles University in Prague; [2]Russian Research University "Higher School of Economics"

This paper is a pilot comparative study on coreference chaining in three languages, namely, Czech, English and Russian. We have analyzed 16 parallel English-Czech newspaper texts and 16 texts in Russian (similar to the English-Czech ones in length and topics). Our motivation was to find out what the linguistic structure of coreference chains in different languages is and what types of distinctions we should take into account for advancing the development of systems for coreference resolution. Taking into account theoretical approaches to the phenomenon of coreference we based our research on the following assumption: the recognition of coreference links for different structural types of noun phrases is regulated by different language mechanisms. The other starting point was that different languages allow pronominal chaining of different length and that coreference chains properties differ for the languages with different strategies for zero anaphora and different systems for definiteness marking. This work reports our first findings within the task of the structural NP types' distribution comparison in three languages under analysis.

DISCOURSE STRUCTURE: A PERSPECTIVE FROM MULTIMODAL LINGUISTICS

Nikolaeva Y. V. (julianikk@gmail.com), **Fedorova O. V.** (olga.fedorova@msu.ru), **Kibrik A. A.** (aakibrik@gmail.com), Institute of Linguistics, Russian Academy of Sciences and Lomonosov Moscow State University, Moscow, Russia

This paper is a step towards multimodal linguistics, considering the verbal form of spoken discourse along with prosodic and gestural phenomena, involved in the process of spoken communication. It is well established that spoken discourse is structured with the help of prosodic features. The basic

segment of talk is elementary discourse unit (EDU), defined on the basis of a set of prosodic criteria and correlated with the semantico-syntactic unit known as clause. A hierarchically more complex unit is sentence. Sentence boundaries are also identified by prosodic features. Illustrative gestures can signal EDU combination into sentences, too. This is performed by gesture assimilation in formal (location of gesture, hand configuration, trajectory, and direction of movement) and content-related (referents, place and time of event) characteristics. One kind of gesture assimilation, catchment, correlates with spoken sentence, whereas the other kind, gestural inertia, with a higher level unit, namely episode. We thus observe partial correlation between the components of multimodal discourse.

VERBS *BYT'* AND *BYVAT'*: CONTEMPORARY STATE AND HISTORY

Paducheva E. V. (elena.paducheva@yandex.ru), Sholokhov Moscow State University for the Humanities, Moscow, Russia

The verb BYVAT' 'to be <iteratively>' (formed with the help of an iterative suffix *-yva-* from *byt'*) belongs to the class of verbs of the **iterative Aktionsart**, which includes such verbs as *xazhivat', slyxivat'* related to the imperfective *xodit', slyshat'*. But BYVAT' occupies a special place in that class. In particular, it is the only one to have an analytical form of the **future tense**. It is claimed that this form exists only in the context of the **motional meaning** of BYVAT' , cf. acceptable *Ja budu byvat' u vas chashche* 'I shall be BYVAT' at your place more often' but **So vremenem ne budet byvat' takix sluchaev* 'over time there won't be BYVAT' in such cases'. The following explanation is given to this fact: motional BYVAT' is included in the aspectual system of Russian not as an **iterative** of the imperfective *byt'* but as an **imperfective** of the momentary perfective *pobyvat'*. No wonder that motional BYVAT' possesses all the properties of an imperfective of a momentary verb: a complete set of tense forms, iterative and general factual meaning of aspect, etc.

The relationship is considered between the verb BYVAT' and the verb *byt'*, which was earlier proved to be perfective in some of its uses and, thus, be-aspectual. The attention is drawn to the fact that BYVAT' is often used as an expressive correlate of *byt'*.

TO BE OR NOT TO BE: CORPORA AS INDICATORS OF (NON-)EXISTENCE

Piperski A. Ch. (apiperski@gmail.com), Russian State University for the Humanities / Russian Academy of National Economy and Public Administration, Moscow, Russia

This paper discusses the notions of acceptability, occurrence, grammaticality and existence, and focuses on the relationship between corpus linguistics and the question of the existence of lexical items. Since corpora are almost exclusively samples from larger populations, it is claimed that they cannot provide evidence for non-existence of words, collocations or constructions. This is because the upper limit of a confidence interval for frequency based on a sample is always greater than zero regardless of the sample frequency. The rule of thumb goes as follows: anything that does not occur in a corpus might have occurred in a similar same-sized corpus zero to five times. If an item occurs in a corpus, this fact can serve as a proof of its existence in the language, but the final decision depends on whether the relevant contexts from the corpus are judged representative of the language variety of interest. In conclusion, I claim that a corpus-based study cannot prove the non-existence of a linguistic item, although it can be used to prove its existence. However, the latter type of proof includes assessing the representativeness of a corpus, which might lead to subjectivity and value judgments.

"I NE DRUG, I NE VRAG, A TAK…": DISTRIBUTION AND PROSODY OF DISCOURSE MARKERS THAT SIGNAL IRRELEVANCE (EVIDENCE FROM THE MULTIMODAL SUBCORPUS OF THE RUSSIAN NATIONAL CORPUS)

Podlesskaya V. I. (podlesskaya@ocrus.ru), Russian State University for the Humanities; Russian Academy of National Economy and Public Administration, Moscow, Russia

The paper focuses on three Russian discourse markers *tak*, *prosto* and *prosto tak*, which fall under a broad category of what is called "loose uses" of language or "vague reference". These are lexical, grammatical and prosodic resources that allow the speaker to refer to objects and events for which the speaker fails to retrieve the exact name, or simply finds the exact name to be unnecessary or inappropriate. The examined discourse markers are employed to signal that

the actual state of affairs is less relevant than another (overtly mentioned or implied) one. The three markers are shown to be associated with different information, syntactic and prosodic structures (e. g. pitch movements). The provided qualitative and quantitative analysis is based on data from the multimodal subcorpus of the Russian National corpus.

LEARNING BY ANALOGY IN A HYBRID ONTOLOGICAL NETWORK

Ponomarev S. V. (serv@newmail.ru), Sputnik LLC, Moscow, Russia

This article describes the general principles of question-answering (QA) system, which produces answers to questions by analogy with the answers and the questions at training sets. As a knowledge base the system uses a number of ontological information of words and expressions from open-access sources and statistic information, collected by processing large text corpora.

The knowledge base is presented as a hybrid ontological network—an oriented graph, where vertices are the words and expressions and edges are the links between words. In addition, each link between two words or expressions is oriented, typified and weighted. The link type characterizes the information source, from which this link and its type were extracted (for example, synonym from Wiktionary). Link weight is determined by reliable information source. All links, obtained from dictionaries and ontological bases, have the weight equals to one. The links, collected by processing text corpora, have the weight equals to frequency of relevant agreed bigrams (for example, a bigram adjective + noun).

The structure of the hybrid ontological network characterizes by a large number of links between the network vertices. Besides direct links connecting two particular network vertices, there could be used composite links, passes through intermediate vertices, which leads to cardinally increasing of number of possible ways between vertices.

Here's a training algorithm that allows setting in the hybrid ontological network the links between words and items in term of combinations of weighted paths between network vertices.

ACQUIRING RELEVANT CONTEXT EXAMPLES FOR A TRANSLATION DICTIONARY

Protopopova E. (rhubarb@yandex-team.ru), **Antonova A.** (antonova@yandex-team.ru), **Misyurev A.** (misyurev@yandex-team.ru), Yandex, Moscow, Russia

This paper addresses the problem of automatic acquisition of parallel context examples for a translation dictionary. We extract them automatically from a parallel corpus, relying on word alignments and parse trees. The ranking of the extracted examples is an essential problem, since we need to select the most distinctive and informative contexts. We propose a machine learning approach as an alternative to simple ranking criteria, such as frequency, or mutual information. We perform the analysis of common sources of inadequate context examples and design a set of features, which can possibly distinguish the bad examples from the good ones. We also experiment with vector models (word2vec) in order to get features that are sensitive to semantics. The evaluation result show that the best of our ranking methods yields 31% improvement in accuracy compared to the ranking by frequency, and 20% improvement over the ranking by mutual information. Using vector models also improves the classification performance.

INFORMATION EXTRACTION FROM CLINICAL TEXTS IN RUSSIAN

Shelmanov A. O. (shelmanov@isa.ru), **Smirnov I. V.** (ivs@isa.ru), Institute for Systems Analysis of Russian Academy of Sciences, Moscow, Russia; **Vishneva E. A.** (vishneva@nczd.ru), Scientific Centre of Children Health, Moscow, Russia

We present and evaluate the pipeline for processing of clinical notes in Russian. The paper addresses the tasks of drug identification and disease template filling, which are related to entity recognition and relation extraction. The disease template filling consists in recognition of disease mentions in text, mapping them to concepts of a thesaurus, and discovering their attributes. Discovering attributes means identifying corresponding spans in text, linking them to diseases, and normalizing them i.e. determining their generalized meaning from a predefined set. We implemented tools for determining the following attributes of disease mentions: negation; the flag indicating the disease mention is not related to a patient; severity; course; and body site. For different tasks, we used dif-

ferent techniques: rule-based patterns and several supervised machine-learning methods. Since there were no annotated corpora of clinical notes in the Russian language available for research purposes, we annotated a dataset, which we used for training and evaluation of the developed tools. The created corpus is available for researchers through the data use agreement.

ON SUMMARIZATION SUPPORTING READABILITY AND TRANSLATABILITY

Sheremetyeva S. O. (lanaconsult@mail.dk), LanA Consulting ApS, Copenhagen, Denmark
National Research South Ural State University, Chelyabinsk, Russia

The article describes a methodology of developing an interactive computer system for supporting a single document text-to-text summarization process focusing on providing for high readability and translatability of the generated summary that, in turn, facilitates further human or automatic processing of the summary text, translation being the most important. The decisions on content selection is delegated to a human but are largely supported by the system. High readability and translatability of the generated text is provided by controlling the syntax of the nascent summary. The approach is a combination of empirical and rational NLP techniques and incorporates a language independent algorithm and language-dependent knowledge base. The validity of the approach was proved by its implementation into a summarizer for scientific papers in the domain of mathematical modelling in the Russian language. The summarizer is fully operational. The methodology presented in this paper is highly portable and allows for extending the summarizer to other domains and languages.

RUSSIAN LANGUAGE-SPECIFIC LEXICAL UNITS IN PARALLEL CORPORA: PROSPECTS OF INVESTIGATION AND "PITFALLS"

Shmelev A. D. (shmelev.alexei@gmail.com), Moscow Pedagogical State University, Moscow, Russia; Vinogradov Institute of Russian Language, Russian Academy of Sciences, Moscow, Russia

The paper deals with language-specific lexical units as they appear in parallel corpora and the degrees of linguistic specificity. It discusses new insights into the languages compared that parallel corpora can provide as well as various pitfalls on the way to an accurate account of typological and cultural differences and similarities.

In particular, it deals with Russian language-specific words, which defy translation into other languages. On the other hand, Russian language-specific words quite often appear in translations into Russian even though no exact equivalent exists in the language of the original text; special attention is given to particles. The lack of such a particle where the communicative situation calls for it every so often gives the impression that we deal with a word-for-word translation of the original text containing no similar marker. In the absence of the relevant particle, the wrong implicatures may appear, or the text may cease to have coherence, or the utterance is perceived as a manifestation of arch use of language.

NATURAL LANGUAGE GENERATION, PARAPHRASING AND SUMMARIZATION OF USER REVIEWS WITH RECURRENT NEURAL NETWORKS

Tarasov D. S. (dtarasov3@gmail.com), ReviewDot Research, Kazan, Russia

Multi-Document summarization and sentence generation are important challenges in natural language processing. This paper presents recurrent neural network (RNN) architecture capable of producing abstractive document summaries, as well as generating novel paraphrases of input sentences in the same language. We demonstrate practical application of our system on the task of multiple consumer reviews summarization.

CONJUNCTIONS *I* 'AND' VS. *NO/A* 'BUT' BETWEEN TWO COORDINATE CLAUSES (REFINEMENT OF THE TERMS "EXPECTATION" AND "NORM")

Uryson E. V. (uryson@gmail.com), Russian Language Institute, Russain Academy of Sciences, Moscow, Russia

The paper deals with the Russian coordinating conjunctions *i* 'and' vs. *no/a* 'but' in a compound sentence. It is common knowledge that in a sentence like "*Q, i P*" the conjunction *i* 'and' marks correspondence to a certain "norm" while in a sentence like "*Q no/a P*" an adversative conjunc-

tion 'but' marks discrepancy between "norm" and the expressed state of affairs. The problem is that in some cases both *i* 'and' and *no/a* 'but' can be used. Another problem is that in some cases the usage of these conjunctions hardly can be interpreted in the terms of "norm". I demonstrate that relevant facts can be adequately described if the basic concept of semantic interpretation is "expectation", but not "norm". Expectation can be induced by (a) common knowledge of laws of nature; (b) common notions of human life, social relations, etc.; (c) text grammar. The conjunction *i* 'and' marks correspondence to the expectation and the conjunctions *no/a* 'but' mark the cancelled expectation in all cases. But the cancelled expectation is obligatorily marked in the case (a), but not (b). As for case (c), text grammar induces merely two expectations: (c1) P has the same general "microtopic" as Q; (c2) P has the same object in the focus as Q. The propositions P and Q can have the same general "microtopic", but different objects in the focus. It is Speaker who chooses strategy for marking this or that case.

RUSSIAN THESAURI AS LINKED OPEN DATA

Ustalov D. A. (dmitry.ustalov@urfu.ru), Ural Federal University, Yekaterinburg, Russia NLPub, Yekaterinburg, Russia

Open linguistic data is a good recently established trend allowing both researchers and developers in the field of natural language processing to create their own applications using high-quality dictionaries, thesauri, corpora, etc. At the same time, the published open data are stored in different formats making them difficult to be used in an efficient way without falling within vendor lock-in. This paper is devoted to the problem of representing popular lexical resources of the Russian language in the form of Linked Open Data. It summarizes the recent work in the field of thesauri representation formats and approaches to converting such formats to those of Linked Data. It also proposes an approach to converting popular Russian thesauri to the vocabularies that are the essential parts of the Linguistic Linked Open Data Cloud. The proposed approach has been implemented in open source software and the resulted dataset has been made publicly available on NLPub in the Turtle format under the terms of a Creative Commons license.

ARTICLE MEANS ARTICLE: ON ONE PATTERN OF TAUTOLOGIES IN RUSSIAN

Vilinbakhova E. L. (elenavilinb@yandex.ru), St. Petersburg State University, St. Petersburg, Russia

The study based on Internet and corpus data [RNC] deals with a special pattern of tautological constructions in Russian called metalinguistic tautologies. The notion was briefly introduced in 1996 by E. Miki as a label for a set of quite heterogeneous examples, but was not further developed. While other tautologies describe entities in the real world, metalinguistic tautologies refer to the use of a linguistic expression. Such constructions show that the speaker is employing a word or an expression in its common, straight meaning. Therefore, they are most often used when context allows other possible interpretations of the linguistic expressions (such as euphemisation, irony, or hyperbole), and sometimes such alternatives are explicitly spelled out: *Inexpensive means inexpensive, not poor quality*. Metalinguistic tautologies are established in Russian with patterns *X znachit X*, *X oznachaet X* 'X means X', *X eto X* 'X is X', and are distinguished from homonymous constructions by their semantic and pragmatic features.

THE METAPHOR OF CAUSED MOTION IN THE AIR IN THE TARGET DOMAIN OF SPEECH ACT VERBS IN RUSSIAN FROM A CONTRASTIVE PERSPECTIVE

Yakovleva I. V. (irinadubrov@gmail.com), Ulyanovsk State University, Ulyanovsk, Russia

The study is devoted to the metaphor of caused motion in the air in the semantic field of speech act verbs in Russian from a contrastive perspective. This semantic shift is typical for verbs of contactless motion implying contact only in the initial phase like *brosat'* ('to throw'). In Russian verbs implying an action performed with hands demonstrate the outstanding capability of accomplishing the semantic shift under study. Another source domain including predicates implying an ejection of a certain object through a kind of hole, mainly a mouth, is not very productive in Russian and is restricted with some negative connotations. The verbs implying emission of a kind of substance with the whole surface do not normally accomplish this type of semantic shift in Russian and we distinguish this hypothetic source domain only when we study the Russian data against the typological background. The semantic shift under study is subject to the aspectual restrictions brought about by the co-occurrence of the basic speech act verbs *govorit'* ('to speak') and *skazat'* ('to say').

CONTRASTIVE ANALYSIS OF PROSODY: ODESSA REGIONAL RUSSIAN VS. STANDARD RUSSIAN

Yanko T. E. (tanya_yanko@list.ru), Institute of lingustics, Moscow, Russia

There is a considerable body of literature referring to segmental phonetic, syntactical, lexical, and stylistic parameters of Odessa accent. However, the prosodic peculiarities of spoken Odessa Russian are seriously underestimated, particularly if we take into consideration the role Odessa language plays in the Russian culture. This paper is aimed at contrasting the Odessa prosody to the prosody of the standard Russian spoken language. The Russian standard prosody, as it is viewed here, is the system corresponding to the inventory of intonational constructions recognized by E. A. Bryzgunova, including their functions in the standard Russian spoken discourse. Prosody is thus referred to not only as a system of distinctive features of the spoken language but also as the basic means of manifesting the communicative meanings: the illocutionary force, the contrast, the discourse continuity. Prosody is analyzed at the level of distinctive features of pitch accents (such as the pitch movement patterns, patterns of the pitch alignment with the text, intensity), at the level of integral pitch accents as they are represented by E. A. Bryzgunova, and at the level of the pitch accents as manifestations of the communicative meanings. For investigation, a minor working corpus of Odessa speech recordings was set up. The corpus consists of interviews with the speakers of Odessa Russian, short stories about Odessa told by the citizens, cooking recipes, jokes, and funny stories. The software programs Praat and Speech Analyzer were used in the process of analyzing the sounding data. The results presented here are exemplified by frequency and intensity tracings of records from Odessa speech oral corpus.

SET PHRASES: A VIEW THROUGH CORPORA

Zakharov V. P. (vz1311@yandex.ru), Saint-Petersburg State University, Saint-Petersburg, Russia

The study of word collocatibility is one of the main tasks of linguistics. Syntagmatic relations bind together language units being in direct contact with each other. The combinatory ability of language units, collocatibility, is one of the linguistic syntagmatic laws. This phenomenon is the main object of the phraseology and lexicography. The article deals with set phrases of different types from the point of view of their numerical evaluation. Corpus linguistics understand set phrases as statistically determined unities. This approach is the basic point of different automatic ways to extract idioms and collocations. The paper describes experiments which show how text corpora and corpus methods and tools such as association measures, word sketches, concordances can be used to expand the entries in existing dictionaries and how set phrases could be evaluated quantitatively. There are a small numbers of works on set phrases productivity during time periods because of small size of historical corpora. In this research examined set phrases usage was studied diachronically on the base of the big Google books Ngram Viewer Russian corpus counting billions of tokens. The study argues that diachronic productivity is best evaluated with a studying contexts. Used corpus tools enable to do it. Ultimately, it is shown and maintained that corpus linguistics methods and tools allow to create dictionaries of new type which have to include a larger amount of set phrases and collocations than before.

RUSSIAN LANGUAGE-SPECIFIC WORDS AS AN OBJECT OF CONTRASTIVE CORPUS ANALYSIS

Zalizniak Anna A. (anna.zalizniak@gmail.com), IL RAS, Moscow, Russia

The paper summarizes methodological principles and some preliminary results of a project "Contrastive corpus-based study of the specific features of Russian semantic system" currently conducted by a research group on the basis of an aligned Russian-French and French-Russian parallel corpus. The purpose of the research project is to verify, by means of contrastive corpus-based analysis, a number of hypotheses concerning Russian "language-specific" words formulated in the course of previous investigations. We assume that translation equivalents of a language unit in another language can be considered as a source of information about the semantics of the latter. Such approach is particularly efficient in case of language-specific words that usually do not have full-fledged equivalents in other languages. Indeed, there are at least

three possible types of mismatch: a certain semantic component is lacking in the translation equivalent; the translation equivalent includes an additional component, which is absent in the original unit; a certain semantic component is rendered by supplementary means. Each of these types of mismatch provides us with linguistic information that contributes to clarify the semantics of the source language unit and thus to verify the hypothesis of its language-specific status.

APPROACHING V2: VERB SECOND AND VERB MOVEMENT

Anton Zimmerling (fagraey64@hotmail.com), Institute for Modern Linguistic Research SMSUH; Institute of Linguistics, Russian Academy of Sciences, Moscow, Russia; **Ekaterina Lyutikova** (lyutikova2008@gmail.com), Lomonosov Moscow State University; Institute for Modern Linguistic Research SMSUH, Moscow, Russia

The paper discusses constituting properties of V2 languages in a perspective of parametric typology. V2 languages are a small group of syntactically uniform languages sharing a number of parameters constraining the clausal architecture and the finite verb placement. We argue that whereas the generative procedure of deriving V2 by verb movement and feature composition of the target head is correct and has empirical validation, the broader definitions of V2 phenomena found in the contemporary work on the subject that loosen the diagnostic criteria on the single preverbal constituent are counterproductive. So called 'partial' or 'residual' V2 languages, where verb movement to the left-peripheral position is allegedly characteristic for a part of root declaratives, do not exist; at the same time, the verb movement by itself is not sufficient to produce the classic V2 profile.

Авторский указатель

Author's Index

Научное издание

Компьютерная лингвистика и интеллектуальные технологии

По материалам ежегодной
Международной конференции «Диалог»

Выпуск 14 (21). 2015

Том 1. Основная программа конференции

Ответственный за выпуск **А. А. Белкина**
Вёрстка **К. А. Климентовский**

Подписано в печать 08.05.2015
Формат 152 × 235
Бумага офсетная
Тираж 250 экз. Заказ № 52

Издательский центр «Российский
государственный гуманитарный университет»
125993, Москва, Миусская пл., д. 6
Тел.: +7 499 973 42 06

Отпечатано с готового оригинал-макета в типографии
ООО «Издательско-полиграфический центр Маска»
117246, Москва, Научный пр-д, д. 20, стр. 9